Part III

Theories and models of speech perception and production
Introduction
The purpose of this review is to summarize some of the literature on the acoustics and perception of vowels. In spite of the relative youth of this area of study, and of experimental phonetics more generally, the literature on the acoustics and perception of vowels is quite extensive. As a consequence, this review will be selective, and an attempt will be made to focus the discussion on those areas that seem to the author to be the most critical to understanding how vowels are spoken and recognized.

Source-filter theory and vowel production
Source-filter theory has served as the primary model that is used to understand the acoustics of speech production – both vowels and consonants – for a great many years. Figure 9.1 shows a simplified version of source-filter theory for the production of three monotone vowels with a static vocal tract posture. (Not shown in this simplified model is the radiation characteristic, a high-pass filter with a slope of about +6 dB per octave that accounts for the greater efficiency with which higher-frequency components are radiated from the enclosed vocal tract to the outside world.) In this figure, the same glottal source signal serves as the input to three different vocal tract filters whose shapes are controlled by the positions of the tongue, lips, and jaw. The peaks in the vocal tract frequency response curve (FRC) are responsible for the peaks that are seen in the envelope of the output spectrum.

What is a formant?
The term “formant” is the most widely used descriptive and explanatory term in acoustic and perceptual phonetics. Consequently, one would think that the term would have a clearly agreed upon meaning, and that any potential for misunderstanding the meaning of the term would be routinely addressed in introductory texts, but this is not the case. The central issue is that formant is sometimes used to refer to a peak in the envelope of an output spectrum, and the same term is sometimes used to refer to a peak in a vocal tract frequency response.
Figure 9.1 Slightly simplified version of source-filter theory for three static, phonated vowels. A single periodic glottal source signal serves as the input to three different vocal tract filters whose frequency response curves are controlled by the positions of the tongue, jaw, and lips. Amplitudes in the output spectrum are derived by multiplying the amplitudes at each frequency by the gain of the filter at those frequencies.
Figure 9.2  Formant-synthesized vowels with identical vocal tract resonances but different fundamental frequencies. The first three formant frequencies were set to 250, 2100, and 2400 Hz; resonator gains and bandwidths are identical for the two signals. The spectrum envelopes were computed by interpolating between adjacent harmonics (Paul, 1981), followed by smoothing with a 50 Hz Gaussian-weighted running average. Note that neither of the $F_1$ envelope peaks is especially close to the 250 Hz first resonance of the formant synthesizer. Further, the two envelope peaks differ from one another by 78 Hz, or a factor of 1.36. Also note the merger of $F_1$ and $F_2$ into a single envelope peak for the vowel with a 300 Hz $f_0$. 
curve (quite often by the same writer). Fant (1960) explicitly defines formants as “the spectral peaks of the sound spectrum” (p. 20); however, later in the same book Fant appears to use the same term to refer to an FRC peak (e.g., p. 53). The view of a formant as a spectral peak is expressed in other widely read sources, such as Peterson and Barney (1952) and Flanagan (1972), who expresses the idea with exceptional clarity: “The manifestations of these normal modes as spectral peaks in the output sound are called formants” (p. 59) [emphasis added]. On the other hand, Stevens (2000) associates the term formant explicitly with the FRC rather than the output spectrum. Consistent with Stevens’ view is the nearly universal use of the term “formant synthesis” to describe synthesis methods such as Klatt (1980; see also Klatt and Klatt, 1990). Implicit in formant synthesis is the view that a formant is a property of a filter since parameters such as $F_1$, $F_2$, etc., control the resonant frequencies of the filter. Also controlled are parameters such as $A_1$, $A_2$, etc. Although these parameters are referred to as formant amplitudes in Klatt and Klatt, the parameters that are being controlled in the synthesis settings are unambiguously resonator gains, not output amplitudes.

Peaks in FRCs and peaks in output spectra correspond closely in many cases, particularly for the three idealized vowels whose FRCs and output spectra are shown in Figure 9.1. Nevertheless, a peak in an FRC and a peak in the envelope of an output spectrum are not and cannot be the same thing. A simple example can be seen in Figure 9.2, which shows harmonic spectra and spectrum envelopes for two formant-synthesized vowels with identical settings of the formant-frequency control parameters (250, 2100, and 2400 Hz) but different fundamental frequencies (100 vs. 300 Hz). Envelopes of the output spectra were computed by interpolating between adjacent harmonics (Paul, 1981; Hillenbrand et al., 2006), followed by light smoothing. Note that the $F_1$ envelope peaks are not especially close to the 250 Hz setting of the first resonance of the synthesizer for either of the vowels. Further, the two envelope peaks corresponding to $F_1$ differ from one another by 78 Hz, or a factor of 1.36, despite identical resonator frequencies, gains, and bandwidths. Also note the merger of $F_2$ and $F_3$ into a single envelope peak for the vowel with a 300 Hz $f_0$. There is nothing special or contrived about the example in Figure 9.2; many others could be illustrated. The main point is that relationships between FRC peaks and peaks in the output spectrum are not always simple; as a consequence, the use of a single term to apply to two related but quite different concepts is far from ideal. Despite this, any change in the use or definition of the term “formant” at this late stage would appear to be out of the question. The term has been used to refer to two quite different phenomena for too long (but see Titze et al., 2015, for some history, discussion, and proposed solutions).

Some basic facts: Acoustic measurements and intelligibility

In what is almost certainly the most widely cited paper in the area, Peterson and Barney (1952; hereafter PB) measured the acoustic properties ($f_0$ and $F_1$–$F_3$) and the intelligibility of ten nominally monophthongal American English (AE) vowels (/i/, /ɪ/, /ɛ/, /æ/, /ɑ/, /ɔ/, /ʊ/, /u/, /ʌ/, and /ɚ/) spoken by a large group of talkers. While these kinds of measurements had been reported in previous studies, the scale of PB’s work was unprecedented, and would remain so for several decades. Audio recordings were made of two repetitions each of the ten vowel types in /hVd/ words spoken by 76 talkers (33 men, 28 women, and 15 children). Relatively little is said about the dialect of the talkers, except that the women and children were raised primarily in the Mid-Atlantic area, while the dialect background of the men was more heterogeneous.
The 1,520 utterances (76 talkers × 10 vowels × 2 repetitions) were presented for identification in randomized order to 70 men and women, including 32 of the 76 talkers. The regional distribution of the listeners was said to be similar to that of the talkers. Identification results were relatively simple, at least in general terms. The average identification rate was 94.4%, with extremely low error rates for some vowels (e.g., /i/: 0.1%, /ɻ/: 0.3%, /u/: 0.8%). The only vowel type with a double-digit error rate was /ɑ/ (13%), confused mainly with /ɔ/. The primary finding, then, is that the utterances were overwhelmingly identified as the vowel that was intended by the talker, leaving open the obvious question of what information listeners are using to identify the vowels.

Formants and $f_0$ were measured at a single time slice that was judged by visual inspection of broadband spectrograms to be steady-state (“following the influence of the [h] and preceding the influence of the [d],” p. 177). The acoustic measurements were made from narrow band amplitude sections. Formant frequencies and amplitudes were estimated by calculating a weighted average using the frequencies and amplitudes of the harmonics comprising the formant peaks (Potter and Steinberg, 1950). A scatterplot of $F_1$ and $F_2$ is shown in Figure 9.3. It is clear from this familiar figure that there is a strong relationship between the $F_1$–$F_2$ values and the intended vowel category.

Figure 9.4 shows a scatterplot of $F_1$–$F_2$ values from a more recent study by Hillenbrand et al. (1995, hereafter H95). Recordings were made of 12 vowel types (/e/ and /o/ were added to the ten vowel types recorded by PB) in /hVd/ environment from 139 talkers (45 men, 48 women, and 46 10- to 12-year-old children). The general approach of the study was quite similar to PB, with the major departures consisting of: (a) The great majority of the talkers (87%) were from Southern Michigan, with most of the remaining talkers from other areas of the inland north that have been characterized by Labov et al. (1972) as part of the "Northern Figure 9.3 Formant frequency measurements at steady-state from Peterson and Barney (1952). The data have been thinned of redundant data points.

Source: Redrawn from Peterson and Barney (1952).
Figure 9.4  Formant frequency measurements at steady-state from Hillenbrand et al. (1995). The data have been thinned of redundant data points.

Figure 9.4 (Continued)

Cities dialect group; and (b) measurements were made of the contours of $f_0$ and $F_1–F_3$ (using linear predictive coding, analysis) throughout the course of the vowel. To allow comparisons to PB, estimates of steady-state times were made by visual inspection of broadband spectrograms. While the formant values for some of the Michigan vowels differ from those of PB (Figure 9.5), mainly because of dialect differences, the two main features of the PB formant
data are also observed in Figure 9.4: (a) There is a clear relationship between vowel category and the values of $F_1$ and $F_2$, but (b) there is a good deal of within-category variability and overlap among adjacent vowel types (see especially the overlap between /ɪ/-/e/ and /æ/-/ɛ/). In both PB and H95, some but by no means all of this variability is related to differences in vocal-tract length (Figure 9.6). Finally, as with PB, the utterances are overwhelmingly

![Figure 9.6](image-url)  
**Figure 9.6** Formant frequency measurements for men, women, and children (Peterson and Barney, 1952). Note that, on average, formant frequencies increase from men to women to children. Some but not all of the within-vowel-category variability is clearly related to differences in vocal-tract length.
recognized as the vowel intended by the talker, with intelligibility that is very slightly higher (95.4%) than PB despite the addition of two vowel categories (/e/ and /o/).

In light of the excellent intelligibility of the vowels in both of these large-scale studies, it is clear that listeners must be identifying these speech sounds based on information in addition to (or perhaps in place of) $F_1$ and $F_2$. Some of these possibilities will be explored in later sections of the chapter.

Formants, spectral peaks, and spectral shape

An implicit assumption underlying experimental work such as PB is the idea that vowel identity is controlled not by the detailed shape of the spectrum but by the distribution of the two or three lowest formant frequencies. This assumption is so pervasive that it is only occasionally explicitly defended or even described. The alternative to formant theory – typically called a whole spectrum or spectral shape approach – presumes that vowel identity is determined by the gross shape of spectrum envelope. An important point that is sometimes made by advocates of whole spectrum models is that formants only seem to be important because spectral peaks have such a large effect on the overall shape of the spectrum envelope (e.g., Zahorian and Jagharghi, 1993; Rosner and Pickering, 1994).

A significant virtue of formant theory is that formant representations appear to account for a very large number of findings in the phonetic perception literature. A straightforward demonstration of this is the high intelligibility of speech that has been generated from the sparse set of control parameters that are needed to drive a formant synthesizer (e.g., Logan et al., 1989). Further, literature extending well into the 19th century (e.g., Mattingly, 1999) abounds with explanations of a wide variety of acoustic-phonetic phenomena that appeal to the concept of a formant. There are, however, several important problems with formant representations. Bladon (1982) referred to the most important of these as the determinacy problem: Quite simply, after many decades of concerted effort by many capable, imaginative investigators, no one has developed a fully reliable method to measure formant frequencies
automatically from natural speech, even in the absence of challenging conditions such as noise or reverberation (see also Klatt, 1982). As is well known, the central problem is that formant tracking involves more than just extracting envelope peaks; it is necessary to assign each of the peaks to specific formant slots corresponding to \(F_1, F_2, F_3\), etc., while rejecting envelope peaks that arise from sources other than vocal tract resonances.

A closely related problem is that the kinds of errors that are made by listeners do not appear to be consistent with the idea that the auditory system is tracking formants. As Klatt (1982) noted, even the best formant trackers are bound to make errors, and these errors are often large ones since they tend to involve either identifying spurious peaks as formants, or missing formants entirely (e.g., when closely spaced formants merge to form a single peak, as in \(F_1\) and \(F_2\) of /ɔ/, or \(F_2\) and \(F_3\) of /i/ or /ɜ/). Human listeners, on the other hand, nearly always hear a vowel that is very close in phonetic space to the vowel that was intended by the talker, as can be seen in virtually any vowel identification confusion matrix. This fact is more readily accommodated by whole-spectrum models.

A sampling of findings that seem more consistent with a spectral shape model includes the following:

First, a statistical pattern classification study by Zahorian and Jagharghi (1993) showed that naturally spoken vowels in CVC syllables could be recognized with greater accuracy based on spectral shape than formants.

Second, a deceptively simple experiment by Lindqvist and Pauli (1968) showed that phonetic boundaries along a Swedish front-vowel continuum ([i]–[y]–[ʉ]) were unaffected by very large changes (as much as 25 dB at the extremes) in the level of \(F_1\) in relation to \(F_2\) and \(F_3\) (see also Ainsworth and Millar, 1971).

Third, Ito and Yano (2001) created test signals in which either \(F_1\) or \(F_2\) was suppressed while maintaining spectral shape that was as close as possible to the original signals (see figure 1 from Ito and Yano). Listener responses to the formant-suppressed signals were largely similar to those of the original signals, suggesting that formant frequencies are not decisive in determining the vowel percept. (See Kieffte and Kluender, 2005, for related experimental work that led the authors to conclusions that are very different from Ito et al.)

Few whole-spectrum models have been specified as explicitly as Bladon and Lindblom (1981), who hypothesized that listener judgments of vowel quality are based on representations of loudness density versus pitch. Critical-band spectra derived from an auditory model (Schroeder et al., 1979) were used in combination with a metric of auditory-perceptual distance (Plomp, 1970) that could be used to predict the similarity of pairs of loudness-pitch spectra. To compare the model output to perceptual data, listeners heard pairs of signals consisting of four-formant /i/ or /y/ and one of seven versions of two-formant vowels with a fixed \(F_1\) and with \(F_2\) values ranging in third-octave steps between 1 and 4 kHz. Listeners were asked to judge the similarity in quality between different pairs of signals. The model output correlated well with the listener judgments (\(r = 0.89\)). However, in discussing their results, the authors noted that

Our distance metric is overly sensitive to the ordinate dimension of the pseudo-auditory patterns [i.e., loudness]... We evidently ought to look, at least in some cases, for a way of reinstating the spectral peak notion while not discarding the benefits that attach to our whole-spectrum measure.

\(p. 1421\)
The issue of sensitivity to loudness differences was explored in an unusually insightful paper by Klatt (1982), which describes experimental work that was directed precisely at the formants/spectral-shape issue. Listeners were presented with pairs of synthetic signals consisting of a standard /æ/ and one of 13 comparison signals. The comparison signals were generated by manipulating one of two kinds of parameters: (a) those that affect overall spectral shape (e.g., high- or low-pass filtering, spectral tilt, changes in formant amplitudes or bandwidths, the introduction of spectral notches, etc.) – but without changing formant frequencies – or (b) formant frequencies, alone and in combination (e.g., $F_1$, $F_2$, $F_1$ and $F_2$, . . .).

Listeners were asked to judge either (a) the overall psychophysical distance, or (b) the phonetic distance between the standard and comparison stimuli, “ignoring as much as they can any changes that are associated with a change in speaker or recording conditions” (Klatt, 1982, p. 1278). The results (Table 9.1) were quite clear in showing that formant frequencies were easily the most important parameters affecting phonetic quality. Other spectral differences – while readily audible to listeners – did not result in large changes in phonetic quality. The task was repeated with /ɑ/ and with whispered vowels, with very similar results. Klatt also showed that the metric used in Bladon and Lindblom correlated very poorly ($r = 0.14$, p NS) with phonetic distance judgments made from the /ɑ/ test signals. Klatt summarized his findings in this way:

Of the stimulus manipulations included in this study, only formant frequency changes induced large changes in phonetic distance. Even though filtering and spectral tilt conditions produce substantial changes in the spectrum [and in psychophysical distance judgements], these changes are apparently ignored when making phonetic judgements. (p. 1278)

Klatt did not see his findings as favoring formant theory, which he found implausible based on considerations such as those previously discussed. He argued that a spectral distance metric was needed that would be maximally sensitive to differences in spectral peaks.

Table 9.1  Results from Klatt’s (1982) experiment on perceived psychophysical (Psy) and phonetic (Phon) distances between pairs of vowels with a quality similar to /æ/.

<table>
<thead>
<tr>
<th>Spectral Shape Manipulations</th>
<th>Psy</th>
<th>Phon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Random phase</td>
<td>20.0</td>
<td>2.9</td>
</tr>
<tr>
<td>2. High-pass filter</td>
<td>15.8</td>
<td>1.6</td>
</tr>
<tr>
<td>3. Low-pass filter</td>
<td>7.6</td>
<td>2.2</td>
</tr>
<tr>
<td>4. Spectral tilt</td>
<td>9.6</td>
<td>1.5</td>
</tr>
<tr>
<td>5. Overall amplitude</td>
<td>4.5</td>
<td>1.5</td>
</tr>
<tr>
<td>6. Formant bandwidth</td>
<td>3.4</td>
<td>2.4</td>
</tr>
<tr>
<td>7. Spectral notch</td>
<td>1.6</td>
<td>1.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Formant Frequency Manipulations</th>
<th>Psy</th>
<th>Phon</th>
</tr>
</thead>
<tbody>
<tr>
<td>8. $F_1$ ±8%</td>
<td>6.5</td>
<td>6.2</td>
</tr>
<tr>
<td>9. $F_2$ ±8%</td>
<td>5.0</td>
<td>6.1</td>
</tr>
<tr>
<td>10. $F_1$, $F_2$ ±8%</td>
<td>8.2</td>
<td>8.6</td>
</tr>
<tr>
<td>11. $F_1$, $F_2$, $F_3$, $F_4$, $F_5$, ±8%</td>
<td>10.7</td>
<td>8.1</td>
</tr>
</tbody>
</table>
and minimally sensitive to spectral tilt, formant-level differences, and other spectral features 
that have little effect on phonetic distance judgments. Klatt proposed a measure in which, 
at each frequency, spectral slope rather than level differences are computed: “In this way, 
a formant peak with the same frequency location in each spectrum but a difference in peak 
height would have exactly the same (zero) slope at the peak and very similar slopes on the 
adjacent skirts.” (Klatt, 1982, p. 1280). In addition, the slope-difference vector at each frame 
is multiplied by a weighting function such that greater weight is given to slope differences at 
envelope peaks, especially the peak with the highest amplitude. Klatt reported a correlation 
of 0.93 between the weighted slope measure and listener judgments of phonetic distance 
from the /ɑ/ experiment.

While Klatt did not believe that the proposed slope measure was the final word on the 
subject (see especially his Comment #5 under Conclusions), the weight of evidence sug-
gests that neither formant representations nor a psychophysically motivated representation 
of envelope shape are capable of accounting for judgments of phonetic quality: “a metric 
is needed that attends to the locations of prominent energy concentrations, but does not 
attend greatly to their relative intensities, nor to the shape of the spectrum in the valleys 
between energy concentrations” (Klatt, 1982, p. 1281). Another peak-dominated spectral 
shape approach based on exactly this premise, but quite different in implementation, will be 
discussed in the next section.

Vowel recognition directly from harmonic spectra

Common to nearly all spectral shape models is the use of a smooth rather than a narrow 
bond spectrum (i.e., a harmonic spectrum for the common case of phonated vowels) to 
represent the to-be-recognized input signal. In an insightful paper, de Cheveigné and Kawahara (1999) argue against smoothing. The centerpiece of their argument is illustrated in 
Figure 9.7. The dashed line in the top of panels a–c shows the idealized envelope shape 
for the vowel /ɑ/ at fundamental frequencies of 100, 200, and 300 Hz. The idealized enve-
lopes are overlaid on harmonic spectra. As de Cheveigné and Kawahara note, the individual 
harmonics of the voice source effectively sample the idealized envelope shapes at discrete 
frequencies. At the bottom of each panel is the spectrum obtained through cepstral smooth-
ing of the three harmonic spectra. It can be seen that the smoothed spectrum is a good fit to 
the idealized envelope at 100 Hz, only a reasonably good fit at 200 Hz, and a very poor fit at 
the moderately high $f_0$ of just 300 Hz. The authors point out that this occurs at the higher $f_0$s 
because the wide harmonic spacing undersamples the idealized envelope shape, resulting 
in alias-induced distortion in the same way that aliasing occurs in an undersampled audio 
signal.

In response to aliasing and other problems with smoothing, de Cheveigné and Kawahara 
proposed a missing data model of vowel identification in which narrow band rather than 
smoothed spectra serve as the input to the recognition algorithm. Harmonic input spec-
tra are then directly compared to smooth templates for each vowel category, with spectral 
differences computed only at harmonic frequencies (i.e., there is no smoothing of input 
spectra and, consequently, no alias-induced distortion). Simulations were run using syn-
thetic signals for five vowel types produced at $f_0$s between 20 and 300 Hz. Templates for 
each vowel were generated from known envelope shapes based on synthesis parameters. 
Findings showed good recognition rates for the missing data model across wide variation 
in $f_0$. On the other hand, an otherwise identical algorithm using smoothed spectra as inputs 
showed extreme sensitivity to $f_0$. 

230
Figure 9.7  (a) Top: envelope (dotted line) and short-term magnitude spectrum of /a/ at $F_0 = 100$ Hz (full line). Bottom: Smoothed short-term spectrum. Smoothing was performed by taking the Fourier transform of the magnitude spectrum, setting it to zero for lags larger than the Nyquist lag $T_0 = 1/2F_0$ (5 ms), and taking the inverse Fourier transform. The smoothed spectrum consists entirely of components below the Nyquist lag. (b) Same, at $F_0 = 200$ Hz. Note the ripples with a period corresponding to the inverse of the Nyquist lag (2.5 ms) that indicate that aliasing is taking place. (c) Same, at $F_0 = 300$ Hz, with the addition of the smoothed spectral envelope (dotted curve at bottom). The spectral envelope was smoothed by removal of lag components beyond the Nyquist lag. The difference between smoothed envelope and smoothed spectrum is the result of aliasing.

Source: Figure and caption reprinted from de Cheveigné and Kawahara (1999).
A narrow band pattern-matching model developed by Hillenbrand and Houde (2003) has features in common with Chevigné and Kawahara, but the model was also strongly influenced by Klatt’s (1982) findings on the disproportionate influence of spectral peaks in relation to other features of the spectrum in controlling phonetic quality. As with de Chevigné and Kawahara, inputs to the recognition model consist of unsmoothed narrow band spectra that are compared to a set of smooth templates for each vowel category. Figure 9.8 shows the signal processing steps that are used to prepare input spectra, beginning with a 64-ms Fourier spectrum using linear amplitudes. This is followed by spectrum level normalization (SLN): a gain function is computed as the inverse of a 1266 Hz Gaussian-weighted running average. The Fourier spectrum is then multiplied by this gain function, increasing the spectrum level in low energy regions and vice versa, thereby reducing (but not eliminating) variation in formant-level relationships and overall spectral tilt. Next, a masking threshold is calculated as a 328 Hz Gaussian-weighted running average. The masking threshold is then subtracted from the level-normalized spectrum, with values below the masking threshold set to zero. The purpose of the masking operation is to emphasize higher energy components (which lie above the masking threshold), to eliminate or reduce the level of spectral components in between harmonics and in the valleys between formant peaks, and to further reduce variation in spectral tilt. The final step involves scaling the peak amplitude in each spectrum to a constant.

The reference patterns that define each vowel category consist of a sequence of smooth spectra (Figure 9.9). A sequence of spectra is used rather than associating each vowel

Figure 9.8  Signal processing steps used in the narrow band pattern matching model: (a) FFT computed over a 64-ms Hamming-windowed segment and broadband spectrum-level normalization (SLN) function computed as the inverse of a 1266-Hz Gaussian-weighted running average of spectral amplitudes; (b) spectrum after processing by the SLN operation and a masking (i.e., threshold) function computed as a 328-Hz Gaussian-weighted running average of spectral amplitudes; (c) spectrum after masking, i.e., after zeroing all spectral values that lie below the level of the masking function; and (d) amplitude normalization, implemented by scaling the largest peak in the spectrum to a constant.

Source: Redrawn from Hillenbrand and Houde (2003).
category with a single template because of the large body of evidence (discussed in the next section) implicating a significant role for spectral change in vowel perception. The templates were derived empirically simply by averaging the harmonic spectra of like vowels at similar times during the course of the vowel (e.g., 15%, 30%, . . .). Separate templates are computed for men, women, and children based on an analysis of about 40 tokens per vowel category.2

Signals in the H95 database were used to create the templates, excluding the relatively small number of tokens with high identification error rates. Templates from the women’s data for 6 of the 12 vowels sampled at 30% of vowel duration are shown in Figure 9.10. Note the merger of $F_1$ and $F_2$ in the /ɔ/ template and the merger of $F_2$ and $F_3$ in the /ɚ/ and /i/ templates. This means that the templates do not necessarily preserve formants in the traditional sense.

The recognition algorithm for a few vowel categories at a few time points is illustrated in Figure 9.11. The 256-point narrow band spectrum computed at 15% of vowel duration is subtracted point-for-point from each of the 12 vowel templates computed at that same time point (only four of which are shown here); the narrow band input spectrum at 30% of vowel duration (not shown) is then subtracted point-for-point from each the 12 vowel templates computed at 30% of vowel duration, and so on. It is important to note that spectral


Figure 9.9 Sequence of five spectral shape templates for /æ/ computed at 15%, 30%, 45%, 60%, and 75% of vowel duration. The templates were derived by averaging the harmonic spectra of like vowels at similar times during the course of the vowel (e.g., 15%, 30%, . . .) – with each spectrum being processed using the steps that are illustrated in Figure 9.8 (i.e., SLN, masking, and peak amplitude normalization). Averaging is followed by light smoothing. Separate templates are computed for men, women, and children based on an analysis of about 40 tokens per vowel category. The templates shown in this figure are from men. Note that successive templates have been offset on the amplitude scale so that the change in spectral shape over time can be seen more clearly.
differences are calculated at all 256 frequencies and not just at harmonic frequencies as in the de Cheveigné and Kawahara model. The vowel that is recognized is the template type (/i/, /ɪ/, . . .) that produces the smallest difference accumulated over the five time points (or two or three, or however many slices are used).

The recognition model was tested using the 139-talker, 1,668-token H95 database. The recognition rate averaged across all utterances was 91.4%, which compares with 95.6% intelligibility for the same signals. Recognition rates were about two percentage points higher for the adults than the children, a pattern that was also shown by the listeners in H95. The confusion matrix resembled that of listeners (e.g., /ɑ/ was confused mainly with /ɔ/ and /ʌ/, /ɛ/ with /ɪ/ and /æ/, etc.), and individual tokens that were incorrectly recognized by the model tended to be the ones that were less intelligible to listeners.

The narrow band model was run (a) without SLN, (b) without masking, and (c) without either SLN or masking. Results were quite clear: Removal of SLN produced a modest drop of about two percentage points, while performance plunged from 91.4% to 59.9% with the removal of the masking operation. The dramatic drop in performance with the removal of masking reflects the importance of high-energy regions of the spectrum that are strongly emphasized through masking. Although the details of the recognition algorithm described here are quite different from Klatt’s (1982) slope metric, both methods emphasize spectral peaks at the expense of other spectral details. It is not clear why the SLN produced such a modest improvement in performance, but the problem may well lie in the details of the implementation rather than the general concept.

In general, the narrow band model findings suggest the following: (a) Naturally spoken vowels can be recognized directly from narrow band spectra (i.e., without first recovering the envelope shape through smoothing); (b) vowels can be recognized accurately, and in
a way that generally resembles the behavior of listeners, without using formants; (c) accurate recognition is possible even when some of the templates do not preserve formants in the usual sense of the term; and (d) as in Klatt’s (1982) study, recognition accuracy improves dramatically when steps are taken to emphasize the role of the spectral peaks that have been shown to have the greatest influence on judgments of phonetic quality. However, these peaks do not necessarily need to correspond to formants in the traditional sense.

**The influence of spectral change on vowel perception**

As discussed earlier, the acoustic measurements made by PB were sampled at a single time point at which the spectrum was judged by eye to be more-or-less stationary. It might be concluded from this measurement approach that investigators of that era made the tacit assumption that all of the information needed to identify a vowel was to be found in a single time slice. This view was summarized nicely by Tiffany (1953):

> It has been commonly assumed or implied that the essential physical specification of a vowel phoneme could be accomplished in terms of its acoustic spectrum as measured over a single fundamental period, or over a short interval including at most a few cycles of the fundamental frequency. That is to say, each vowel has been assumed to have a unique energy vs. frequency distribution, with the significant physical variables all accounted for by an essentially cross-sectional analysis of the vowel’s harmonic composition.

\[^{p. 290}\]
Tiffany, however, questioned this assumption and suggested the possibility that changes in the spectrum throughout the course of a vowel may influence the percept. Similarly, in an early description of the PB vowels, Potter and Steinberg (1950) observed:

It should be noted . . . that we are representing a vowel by a single spectrum taken during a particular small time interval in its duration. Actually, a vowel in the word situation . . . undergoes transitional movements from initial to final consonant. Not only does the spectrum of the vowel change with time, but the ear in identifying the word has the benefit of all of the changes.

(p. 815)

Finally, the following statement is very nearly the last thing that PB say in their paper:

It is present belief that the complex acoustical patterns represented by the words are not adequately represented by a single section, but require a more complex portrayal. The initial and final influences often shown in the bar movements of the spectrogram are of importance here. The evaluation of these changing bar patterns . . . is, of course, a problem of major importance in the study of the fundamental information bearing elements of speech.

(p. 184)

In light of the early awareness of the potential importance of spectral change in characterizing vowels – clearly stated by several investigators in papers that received a good deal of attention – it is surprising that this issue did not receive much attention for many years. But despite the late start, there is now a significant body of evidence (mainly for the North American English vowels that have been most heavily studied) showing that spectral movement plays an important role in vowel perception. Evidence supporting this conclusion comes from several sources, including: (a) Measurement and pattern recognition studies showing that individual North American English (NAE) vowels have distinctive spectral change patterns, and that incorporation of this dynamic information in pattern recognition models significantly improves classification accuracy; (b) listening studies with excised signals showing that stationary targets are not necessary for vowels to be recognized accurately; and (c) listening studies with static vowels showing that stationary targets are not sufficient for accurate vowel recognition.

Spectral change 1: Measurement and pattern recognition

The top panel of Figure 9.12 shows average formant frequencies measured at the beginnings and ends of ten vowel types spoken in isolation by five men and five women from Western Canada (Nearey and Assmann, 1986). The bottom panel of this figure shows similar data from 12 vowel types in /hVd/ syllables spoken by 45 men from southern Michigan and other areas of the inland north (Hillenbrand et al., 1995; data from the H95 women and children show similar patterns). There are clearly some dialect-related differences between the two datasets, but also many similarities. There are a few vowels – mainly /i/ and /u/ – that show very little spectral movement, but most of the vowels show substantial diphthong-like changes over the course of the vowel, patterns that have been referred to as vowel inherent spectral change (VISC; Nearey and Assmann, 1986). The phonetic diphthongs /e/ and /o/ show precisely the kinds of spectral movements that are expected, but the degree
Figure 9.12  Top panel: Formant frequencies measured at the beginnings and ends of ten Western Canadian vowel types spoken in isolation by five men and five women (Nearey and Assmann, 1986). Arrow heads in both panels indicate vowel offsets. Bottom panel: Formant frequencies measured at the beginnings and ends of 12 American English vowel types in /hVd/ syllables (Hillenbrand et al., 1995). The utterances were spoken by 45 men.

Source: The top panel is reprinted from Nearey and Assmann (1986); the bottom panel is redrawn from Hillenbrand et al. (1995).
of spectral change shown for these two sounds is not unusual in relation to several of the nominally monophthongal vowels in both sets of data (see especially /ɪ/, /ɛ/, and /æ/ in the Western Canada data and all but a few of the vowel types in the Michigan data). The obvious possibility that is suggested by these measurements is that individual vowel types in the fairly crowded English vowel space are rendered more distinctive by differences in spectral change patterns. For example, in the Michigan data /æ/ is raised and fronted, creating a high degree of overlap between /æ/ and /ɛ/ when represented in static formant space (Figure 9.4). However, listeners in H95 showed an error rate of just 5% for these vowels, suggesting that the distinctive spectral change patterns (along with duration differences, which will be discussed later) enhance the intelligibility of these vowels. A similar role for spectral change is suggested for /ɪ/ and /e/, another pair of vowels that show a good deal of overlap in static formant space but show spectral change patterns that are quite different from one another (compare Figure 9.4 with the lower panel of Figure 9.12).

Evidence from pattern recognition studies shows conclusively that vowels can be classified with greater accuracy using features sets that incorporate spectral change as compared to otherwise comparable classification models that rely on measurements from a single time slice. For example, Zahorian and Jagharghi (1993) used a statistical pattern recognizer to classify nine vowel types in CVC syllables spoken by ten men, ten women, and ten children in the environment of nine initial consonants and eight final consonants (not fully crossed). The recognition algorithm was driven by features consisting of either formant frequencies or a discrete cosine transform representation of spectral shape. For both formant and spectral-shape features, the authors reported much better classification accuracy, and better agreement with listener data, when the classifier was driven by dynamic rather than static feature vectors. (They also reported better performance with spectral shape than formants.)

Similarly, using various combinations of \( f_0 \) and formants as features (e.g., \( F_1 - F_3 \), \( F_1 - F_2 - F_3 \), \( f_0, F_1 - F_2 \), etc.), H95 reported consistently better classification accuracy when the formant pattern was sampled twice (at 20% and 80% of vowel duration) than otherwise comparable models sampled once at steady-state (Table 9.2). Incorporating a third sample of the feature vectors at 50% of vowel duration produced little or no improvement in classification performance. A very similar finding using a different recognition model was reported by Hillenbrand and Houde (2003). These findings are consistent with the idea that listeners evaluate VISC based primarily on vowel onsets and offsets – the dual-target model proposed by Nearey and Assmann (1986).

A consistent advantage for dynamic rather than static features was also found in Hillenbrand et al. (2001) for formants measured from CVC syllables spoken by six men and six women using all combinations of seven initial consonants, eight vowel types, and six final consonants. As Stevens and House (1963) showed in a study using only symmetrical

<table>
<thead>
<tr>
<th>Parameters</th>
<th>1 sample</th>
<th>2 samples</th>
<th>3 samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F_1, F_2 )</td>
<td>76.1</td>
<td>90.3</td>
<td>90.4</td>
</tr>
<tr>
<td>( F_1 - F_3 )</td>
<td>84.6</td>
<td>92.7</td>
<td>93.1</td>
</tr>
<tr>
<td>( f_0, F_1, F_2 )</td>
<td>82.0</td>
<td>92.5</td>
<td>92.6</td>
</tr>
<tr>
<td>( f_0, F_1 - F_3 )</td>
<td>87.8</td>
<td>94.1</td>
<td>94.8</td>
</tr>
</tbody>
</table>
consonant environments (/bVb/, /dVd/, /gVg/, etc.), there are many statistically reliable effects of consonant environment on vowel formants. In both studies, two of these effects were large (upward shifts in $F_2$ of ~500–700 Hz for /u/ and ~200–250 Hz for /ʊ/ in the environment of initial – but not final – alveolars), but most of the effects were not especially large in absolute terms. It is possible that complications due to these context effects would attenuate or even eliminate the performance advantage for dynamic feature sets. However, in Hillenbrand et al. (2001), pattern classification accuracy for features measured from these heterogeneous syllable types was consistently better with two rather than one sample of the formant pattern. The advantage in pattern classification accuracy, then, is not an artifact of simple, fixed consonant environments.

Spectral change 2: The non-essential nature of static targets

A series of experiments initially conducted at the University of Minnesota used a “silent center” technique to show conclusively that the static targets that might have been considered essential to the vowel percept were simply not needed. Shown in the top row of Figure 9.13, from Jenkins et al. (1983), are schematic representations of the original, naturally spoken /bVb/ syllables, with a shorter vowel on the left and a longer vowel on the right. In the middle row, the center 50% has been excised from the short vowel and the center 65% has been excised from the long vowel, leaving only the onsets and offsets. In the bottom row, the onsets and offsets have been edited out, leaving the center 50% (left) or 65% (right) of the syllable. Results of listening tests were unambiguous: the silent center signals were identified as well (92.4%) as the original utterances (93.1%); i.e., removal of the vowel centers had no measurable effect on vowel intelligibility. The variable centers were also well identified, with no statistically reliable differences among the identification rates for the three conditions.

A follow-up study by Nearey and Assmann (1986) excised 30-ms Hamming-windowed segments from isolated (but not static) vowels at 24% of vowel duration (nucleus) and at 64% of vowel duration (offglide). Listeners were asked to identify: (a) the unedited vowels, (b) the nucleus followed by the offglide, with 10 ms of silence separating the clips, (c) the nucleus segment repeated, and (d) the nucleus and offglide segments played in reverse order. Nearey and Assmann found that excised segments that were presented in natural order were as intelligible as the unedited vowels. However, error rates more than doubled for the repeated-nucleus and reverse-order conditions. The poor intelligibility of the reverse-order signals indicates that the key factor is not the presence vs. absence of spectral change, but rather a pattern of spectral change matching that which is typical of the vowel type that is being uttered. (For additional experimental work on excised vowels – all of it interesting – see Andruski and Nearey, 1992). Jenkins and Strange, 1999; Parker and Diehl, 1984; Strange, 1989; Strange et al., 1994).

Spectral change 3: The insufficiency of static targets

There is convincing evidence from several sources indicating that, for NAE vowels, very high identification rates of the kind reported by PB and H95 cannot be achieved using vowels with static spectral patterns. The most striking demonstration comes from a remarkable study by Fairbanks and Grubb (1961), who went to extraordinary lengths to record the highest quality and most representative examples of the nine static, sustained vowels that they studied (the vowel types studied by PB, excluding /ɚ/). A brief excerpt from a
lengthy set of instructions that was given to their talkers will give the reader some idea of the extraordinary lengths to which the experimenters went to record prototypical instances of each vowel:

Essentially what we are trying to do is to collect samples of each vowel that are as nearly typical or representative of that vowel as possible. More specifically, we are interested in samples that depict the central tendency of each vowel. . . . Another way of putting the problem is to say what we want you to do is to imagine the target on the basis of your experience in listening to speech, and then demonstrate what the target is
by producing a vowel of your own that hits the target as you imagine it. You will understand from this that we are trying to get samples that are something more than merely acceptable and identifiable.

The seven talkers were all men, all phonetically trained, and all were faculty in the Speech and Hearing Department at the University of Illinois. Utterances were immediately auditioned by an experimenter, and talkers often recorded several tokens before an utterance was accepted. The full set of recordings was then auditioned a second time and talkers were asked to return to the lab to re-record any utterances that were judged by the experimenters to be unsatisfactory. The vowels were typically 1–2 s long and were intended to have stationary pitch and timbre; 300-ms segments were then excised from these recordings for presentation to listeners, who consisted of eight phonetically trained graduate students in the same department.

With that lengthy introduction to Fairbanks and Grubb, the results were quite simple: Despite the exacting procedures used by the experimenters to record the highest quality and most representative examples of each utterance, the average intelligibility was just 74.0%—more than 20 percentage points lower than the intelligibility of the /hVd/ utterances recorded by PB and H95, which were spoken by 76 untrained talkers (men, women, and children), and without the elaborate auditioning procedure used by Fairbanks and Grubb. Further, in PB the test signals were identified in a large auditorium by listeners with no phonetics training. Fairbanks and Grubb reported a good deal of variability in the intelligibility of individual vowel types, and these variations are revealing: the highest intelligibility (~91–92%) for these static vowels was seen for /i/ and /u/, vowels that tend to show the least spectral movement (see Figure 9.9). Much lower intelligibility (~53–66%) was reported for vowels that tend to show a good deal of spectral change for speakers in the Inland North (e.g., /ɪ/, /æ/, and /ʌ/). In addition to the static nature of the recordings, the absence of duration variability in the Fairbanks and Grubb vowels almost certainly contributed to the modest intelligibility of the utterances. Experimental work on the role of duration in vowel perception will be discussed in a separate section later in this chapter.

Evidence for the insufficiency of static targets also comes from Hillenbrand and Gayvert (1993), who used a formant synthesizer to generate 300-ms static versions of all 1,520 signals in the PB database using measured values of $f_0$ and $F_1–F_3$. The average intelligibility of these static signals was 72.7%. Once again, this figure is more than 20 percentage points lower than the intelligibility of the original /hVd/ signals, and is quite similar to the value reported by Fairbanks and Grubb for their static vowels. Further, static vowels such as /i/ and /u/ tended to be identified much better (96.2% and 89.1%, respectively) than vowel types that typically show more spectral change. Findings leading to the same basic conclusion were reported by Hillenbrand and Nearey (1999), who asked listeners to identify three versions of 300 /hVd/ syllables drawn from the H95 recordings: (a) a naturally spoken signal (NAT), (b) an original-formant (OF) synthetic version of the same signal generated from the measured formant contours; and (c) a flat-formant (FF) version generated with formants fixed at the values measured at steady-state (Figure 9.14). Results for the critical OF–FF comparison were not subtle: The 88.5% average intelligibility for the OF signals dropped to 73.8% when the formants were flattened. It is striking how similar the intelligibility figures are for the static signals from Fairbanks and Grubb (74.0%), Hillenbrand and Gayvert (72.7%), and Hillenbrand and Nearey (73.8%). Further, as shown by both Fairbanks and Grubb and Hillenbrand and Gayvert, static vowels that typically show relatively little
spectral change are identified better than vowels that typically show more spectral change (Figure 9.15).

Although it is not relevant to the spectral change issue that is being discussed in this section, it is of some interest to note that the naturally spoken signals were significantly more intelligible (95.4%) than the OF formant-synthesized versions (88.5%). This difference of about seven percentage points, which is not trivial, is consistent with Bladon’s (1982) reductionism argument against formant representations, the idea that formant representations discard information that may prove to be phonetically relevant.

In summary, a large body of evidence indicates that NAE vowels are better thought of as distinctive trajectories through acoustic-phonetic space rather than points in that space. In spite of this evidence, accumulated over several decades, the long-standing practice of characterizing vowels as points in acoustic-phonetic space rather than trajectories through that space remains exceedingly common. As we argue in Hillenbrand (2013), these static representations are not so much wrong as they are incomplete.

Figure 9.14 Three stimulus conditions from Hillenbrand and Nearey (1999). NAT: A naturally spoken /hVd/ syllable. OF: A formant-synthesized version of the same signal generated from the original measured formant contours. FF: A formant-synthesized version generated with formants fixed at the values measured at steady-state.

The role of duration in vowel perception

There is good evidence that listeners make use of duration in identifying vowels, although there is less than perfect agreement on the details. The main source of the influence of duration on vowel identification can be seen in Table 9.3, which lists average durations and duration ratios for pairs of adjacent English vowels measured from connected speech by Crystal and House (1988). Generally similar findings have been reported for other kinds of speech material (e.g., Black 1949; van Santen, 1992; Hillenbrand et al., 1995). Most pattern recognition studies have shown better classification accuracy when duration is added to spectral parameter sets. Counterexamples include Zahorian and Jagharghi (1993), who reported non-significant improvements in classification accuracy of less than 1% when duration was added to spectral feature sets consisting of either formant frequencies or spectral shape. Watson and Harrington’s (1999) study of Australian English vowels showed a small improvement in classification accuracy when duration was added to formant measurements, although the duration effect was seen mainly for diphthongs rather than monophthongs. In contrast, H95 reported consistent improvements in the accuracy of a quadratic discriminant classifier in recognizing vowels in /hVd/ syllables when duration measures were added to feature sets consisting of various combinations of $f_0$ and $F_1$–$F_3$. Especially large improvements were seen for the /æ/–/ε/ pair. Similarly, Hillenbrand et al. (2000) reported modest but consistent improvements in classification accuracy with the addition of duration to parameter sets consisting of various combinations of $f_0$ and formant frequencies. The test signals consisted of CVC syllables formed by seven initial consonants, eight vowels, and six final consonants in all combinations spoken by six men and six women.
Evidence that is more convincing than these pattern recognition studies comes from perceptual experiments in which listeners are asked to identify utterances in which vowel duration was directly manipulated. For example, Tiffany (1953) recorded 12 vowel types as long sustained vowels, which were later excised at four durations (80 ms, 200 ms, 500 ms, and 8 s). Several of Tiffany’s findings were consistent with the idea that duration plays a role in vowel identification. For example, vowels with long inherent durations, such as /e/ and /ɑ/, were more likely to be correctly identified when presented at long durations, and vice versa for vowels such as /æ/ and /ʌ/ with short inherent durations. Stevens (1959) synthesized CVC syllables with the vowels /i/, /ɪ/, /ɛ/, /æ/, /u/, /ʊ/, /ʌ/, and /ɑ/ at durations between 25 ms and 400 ms. There were several effects that were consistent with Tiffany’s findings and with the duration data in Table 9.3. For example, at short durations, vowels similar to /æ/ or /ɛ/ were heard more often as /ɛ/, and vowels similar to /ɑ/ or /ʌ/ tended to be heard as /ʌ/. However, the effect of duration on the identification of the /i/–/ɪ/ and /u/–/ʊ/ pairs was much weaker. (See also related work by Ainsworth, 1972, and the somewhat equivocal findings of Huang, 1986).

The experimental work discussed to this point used synthesized vowels that did not incorporate natural VISC patterns. It is now well established that the identity of NAE vowels is not conveyed well to listeners when spectral change patterns are not preserved, suggesting the possibility that the more ambiguous spectral properties of the static vowels that were used in early work may have invited greater reliance on duration, thereby overestimating its influence. Hillenbrand et al. (2000) synthesized 300 /hVd/ signals that were closely modeled on naturally spoken utterances drawn from H95. The synthesizer, which has features in common with McAuley and Quatieri (1986), is similar in general conception to inverse Fourier synthesis, with two major exceptions: (a) Phase relationships are not preserved and, more important, (b) sinusoidal components are synthesized only for (narrow band) spectral peaks, regardless of whether they are harmonic or inharmonic. The intelligibility of signals that are synthesized with this method is quite good, and duration can be artificially shortened or lengthened simply by analyzing the signal at one frame rate and synthesizing the signal at a different frame rate (e.g., duration can be increased by a factor of 1.5 by analyzing a signal with an 8-ms frame rate and synthesizing it at a 12-ms frame rate). When this method is used to alter the duration of a signal, the result is a pair of signals with the same sequence of spectra from one frame to the next, but the rate at which the spectra evolve over time is different from one duration condition to another. The vowels of 300 /hVd/ signals were synthesized: (a) at their original measured duration (OD), within the limits of the 8-ms frame rate that was used; (b) at a short duration (SD) of 144 ms (two standard deviations below

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Original Duration</th>
<th>Short Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>/e/</td>
<td>/ɪ/</td>
<td>1.81</td>
</tr>
<tr>
<td>/i/</td>
<td>/ɪ/</td>
<td>1.59</td>
</tr>
<tr>
<td>/æ/</td>
<td>/ɛ/</td>
<td>1.50</td>
</tr>
<tr>
<td>/u/</td>
<td>/ʊ/</td>
<td>1.48</td>
</tr>
<tr>
<td>/a/</td>
<td>/ʌ/</td>
<td>1.36</td>
</tr>
<tr>
<td>/e/</td>
<td>/ɛ/</td>
<td>1.28</td>
</tr>
<tr>
<td>/ɔ/</td>
<td>/ʌ/</td>
<td>1.06</td>
</tr>
</tbody>
</table>

Table 9.3 Pairs of adjacent American English vowels that differ in inherent duration. Average duration ratios are shown in parentheses. Measurements were made from vowels in stressed syllables in connected speech (Crystal and House, 1988).
Acoustics and perception of English vowels

the grand mean of all vowel durations); (c) at a long duration (LD) of 400 ms (two standard deviations above the grand mean); and (d) at a neutral duration (ND) of 272 ms, the grand mean of all vowel durations. Durations were modified during the vowel only and not during the /h/ and /d/ segments.

Table 9.4 shows the average intelligibility scores from 14 listeners. The 96.0% intelligibility for the OD signals is quite close to the 95.4% value reported by H95 for the full set of naturally spoken /hVd/ signals, indicating that the synthesis method preserves vowel identity. Artificially shortening or lengthening resulted in nearly symmetrical decreases in intelligibility of 4.6 and 5.1 percentage points, respectively. The ND condition, which did not prove to be very revealing, showed a modest drop of 1.9 percentage points.

The ~5% average drop in intelligibility resulting from either shortening or lengthening is not particularly revealing by itself since this figure is arrived at by averaging some cases in which pairs and clusters of vowels were strongly affected by duration with other cases in which duration effects that might have been expected almost never occurred. Confusion matrices were assembled that focused on instances in which the OD version of an utterance was identified correctly, but the SD or LD version was heard as some other vowel (e.g., the OD version might be correctly heard as /æ/ but the SD version as /ɛ/). Table 9.5 summarizes the most frequent labeling shifts resulting from shortening and lengthening. The most common identification shifts resulting from vowel shortening consisted of OD versions of /ɔ/ shifting to either /ɑ/ or /ʌ/, and /æ/ shifting to /ɛ/. As the table shows, the effects of vowel lengthening are nearly the mirror image of those for shortening. These findings, of course, are predictable based on the measurement data from studies such as Crystal and House (1988). However, conspicuously absent from this table are instances of shortened /i/ shifting to /ɪ/ (0.6%) or shortened /u/ shifting to /ʊ/ (0.0%) – in spite of large and reliable differences in typical duration for these pairs. The same was true for lengthened signals; e.g., lengthened /ɪ/ almost never shifted to /i/ (0.1%) and lengthened /ʊ/ rarely shifted to /u/ (1.3%). These findings indicate quite clearly that there is more going on with listeners’ use (and non-use) of duration information than statistical learning based on the magnitude and reliability of measured duration differences in natural speech. As a simple illustration of this point, Crystal and House (1988) reported an inherent duration ratio of just 1.06 for /ɔ/:/ɑ/, but vowel shortening resulted in 41 /ɔ/—>/ɑ/ shifts (14%). On the other hand, there were almost no shifts between /i/ and /ɪ/, despite an inherent duration ratio of 1.59, and no /u/—>/ʊ/ shifts in spite of an inherent duration ratio of 1.48.

An account of these findings was offered suggesting that very little weight is given to duration for pairs such as /i/—>/ɪ/ and /u/—>/ʊ/ because these vowels can be identified with

Table 9.4 Average recognition rates for the four duration conditions (OD = original duration, ND = neutral duration, SD = short duration, LD = long duration). Standard deviations across listeners are shown in parentheses.

<table>
<thead>
<tr>
<th>Duration</th>
<th>Intelligibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>OD</td>
<td>96.0 (2.7)</td>
</tr>
<tr>
<td>SD</td>
<td>91.4 (5.2)</td>
</tr>
<tr>
<td>LD</td>
<td>90.9 (3.3)</td>
</tr>
<tr>
<td>ND</td>
<td>94.1 (2.9)</td>
</tr>
</tbody>
</table>
minimal ambiguity based on spectral information alone. On the other hand, /ɔ/ cannot be as reliably distinguished from /ɑ/ and /ʌ/, or /æ/ from /ɛ/, based solely on spectral information, increasing listeners’ reliance on duration to identify these vowels. To test this idea, a statistical pattern classifier was trained on: (a) \( f_0 \), (b) two samples of \( F_1 \)–\( F_3 \) (at 20% and 80% of vowel duration), and (c) the original, measured duration values. The classifier was then tested using the same spectral measurements but durations that were either short or long by two standard deviations, as simulations of the SD and LD listening conditions. The pattern classifier produced vowel confusions that shared many features with the listener data in Table 9.5, such as large numbers of shifts in recognition involving /ɔ, ɑ, ʌ/ and /æ–ɛ/ but very few shifts involving /i/–/ɪ/ or /u/–/ʊ/.

In summary, duration plays a measurable role in vowel identification. The magnitude of the effect is a modest 5% averaged across all vowels, but the influence of this cue varies quite substantially across different pairs and clusters of vowels. Further, the weight that listeners give to duration appears to depend not only on the magnitude of measured differences in inherent duration for a given vowel pair but also on how accurately individual vowel types can be identified based on spectral information alone.

There is undoubtedly a good deal more to be learned about how listeners make use of duration in identifying vowels. The focus of this work has been on simple utterances such as isolated vowels or citation-form words. The role of duration in connected speech is less certain since it is well known that there is a large set of factors that influence segment duration in conversational speech. As discussed in Klatt’s (1976) comprehensive review, the factors affecting segment duration run the gamut from global dimensions such as speaking rate to word- and phrase-final lengthening to lexical and sentential stress to phonetic details such as the influence of final-consonant voicing on the duration of the preceding vowel. Coming to some understanding of how listeners make use of duration in identifying vowels in the presence of this large and diverse set of competing factors that influence segment duration is a worthy topic for future work on this topic, though it is not likely to be an especially easy problem to address.

**Talker normalization and frequency scaling**

The acoustic characteristics of vowels (and other speech sounds) differ from one talker to the next, even when dialect is held constant, yet listeners typically have little difficulty
recognizing speech across variation in talker characteristics. There is a long history of inquiry into the perceptual and cognitive mechanisms that might be involved in listeners’ accommodation to variation in the characteristics of individual talkers. The literature in this area is vast, so it will be possible in this section to review just a sampling of the work on this problem. Ainsworth (1975) distinguishes intrinsic and extrinsic normalization processes. Intrinsic schemes derive normalizing information entirely from the to-be-recognized token. With extrinsic normalization, calibrating information is derived from other speech sounds uttered by the same talker.

### Intrinsic Normalization

Many intrinsic normalization schemes have been proposed over the years. As Miller (1989) pointed out, many of these models are closely related to the *relative resonance* or *formant ratio* theory proposed by Lloyd in the late 19th century (e.g., Lloyd, 1890, 1892). Lloyd, who based his theory in large part on the central role that is played by frequency ratios in music, proposed that vowels with similar timbres have similar formant ratios. Lloyd’s version of relative resonance theory is appealing in many respects, with the most significant virtue lying in the substantial reduction in variability that formant ratios produce when like vowels are spoken by talkers differing in sex or age (e.g., see Minifie, 1973, Figure 7.15).

Formant ratio theory, however, is not without problems. The most important of these problems is the fact that there are pairs and clusters of vowels with distinct qualities but very similar formant ratios: e.g., /ɑ/–/ɔ/ and /u/–/ʊ/–/æ/.

A particularly influential model proposed by Miller (1984, 1989) uses three spectral distances: \( \log F_3 - \log F_2 \), \( \log F_2 - \log F_1 \), and \( \log F_3 - \log \text{SR} \), where SR (sensory reference) is a transform of the fundamental frequency (\( \text{SR} = 168/(f_0/168)^{1/3} \)). The log formant differences, of course, are fully equivalent to Lloyd’s formant ratios. The virtue of the third spectral difference is that vowel pairs with similar \( F_1/F_3 \) ratios (e.g., /ɑ/–/ɔ/ and /u/–/ʊ/) can be distinguished based on distinct distances between \( F_1 \) and \( F_0 \), addressing the most important problem with Lloyd’s scheme. Miller tested his model using 435 utterances that were taken from several databases. Using hand-drawn “perceptual target zones” that were fit to the data by eye, Miller reported that vowel types could be recognized with 93% accuracy.

Many other versions of formant ratio theory have been proposed over the years, although as Miller (1989) points out, many of these authors do not appear to have been aware of Lloyd’s early work. A model proposed by Syrdal (1985; Syrdal and Gopal, 1986) shares many features with Miller’s model, except that spectral distances are calculated using the Bark scale rather than logs. The three parameters in this model are \( B_3 - B_2 \), \( B_2 - B_1 \), and \( B_1 - B_0 \). Using measurements from PB and a linear discriminant classifier, Syrdal and Gopal reported a recognition rate of 85.7%. While this figure is substantially lower than the 94.5% intelligibility shown by PB’s listeners, at least some of this difference is almost certainly related to the static nature of the PB measurements.

Disner (1980) argued that normalization methods should not only maximize differences between phonetically distinct vowel types, they should also minimize differences in the same vowel type spoken by different talkers. Syrdal (1985) evaluated this criterion by measuring the performance of a linear discriminant model in classifying the PB tokens by talker group (men vs. women vs. children), reasoning that an ideal normalization method would make it difficult to classify tokens of the same vowel type spoken by different talker groups. Syrdal found that the model was very good at classifying tokens by talker group when trained and tested on \( f_0 \) and \( F_1 - F_3 \) in Hz (89.9%), but classification accuracy for talker group
fell to 41.7% when Bark spectral differences were used. The 41.7% correct identification figure for talker group is clearly above the ~33% that would be expected by chance, but Syrdal argued that at least some component of the above-chance performance might reflect dialect differences across the three talker groups (e.g., Byrd, 1992). However, the main point is that the Bark spectral difference representation was much better than absolute linear frequencies at minimizing differences across talker group. A large number of other intrinsic normalization schemes related to the formant-ratio concept have been proposed over the years. Of particular importance are thoughtful schemes described by Peterson (1961) and Nearey (1978, 1992; Nearey et al., 1979).

The work on intrinsic normalization discussed earlier consisted entirely of modeling studies. Modeling studies can be useful in suggesting logically possible perceptual mechanisms, but listening studies are needed to evaluate their validity. Of special interest is the question of whether listener judgments of vowel identity are, in fact, affected by $f_0$. There is no question that $f_0$ exerts a measurable effect on vowel identification, although the details are anything but simple. An early study of static, synthetic vowels by Miller (1953) reported an upward shift in $F_1$ of 80 Hz (16%) on an /ʊ/-/ʌ/ continuum as a result of doubling $f_0$ from 130 to 260 Hz. A much smaller shift of 30 Hz (6%) in the $F_1$ boundary was seen for the same one-octave $f_0$ change on an /ɪ/-/ɛ/ continuum. Among many other results of this kind, Fujisaki and Kawashima (1968) found a 14% $F_1$ shift for a one-octave change in $f_0$ for an /u/-/e/ continuum and a 21% shift for /o/-/ɑ/.

Using a different approach, Ryalls and Lieberman (1982) synthesized static versions of nine vowel types with formants set to the average values for PB’s men and with $f_0$ set to (a) an average value (135 Hz), (b) a lower-than-average value (100 Hz, 5.2 semitones lower than the average value), or (c) a much higher than average value (250 Hz, 10.7 semitones higher than the average value). The authors reported no increase in the identification error rate when $f_0$ was decreased from 135 to 100 Hz, but an $f_0$ increase from 135 to 250 Hz produced significantly more identification errors. It is not entirely clear how this finding should be interpreted since, among other considerations, the increase in pitch from the average value to the high value was substantially greater than the decrease from the average value to the low value. A second experiment used average formant frequencies for PB’s women. For reasons that are not entirely clear, the same three $f_0$ conditions (100, 135, and 250 Hz) were used for the women’s signals as well. Results for these signals, which are not described in much detail, showed a higher error rate for the 250 Hz signals (inferential statistics are not reported for experiment 2). This finding is a bit surprising since the 250 Hz condition is only slightly higher than the 224 Hz average $f_0$ reported for PB’s women. The main conclusion reached by the authors is that the error rate increased for the high $f_0$ condition because formants (or perhaps the envelope shape) are more poorly specified at higher $f_0$s due to the wider harmonic spacing. It is not entirely clear what the findings of this study might have to say about the role of $f_0$ in intrinsic normalization (and that does not appear to have been their purpose). However, for Experiment 1, the failure to find a difference in error rate between the average and low $f_0$ conditions would seem to be at odds with models such as Miller (1989) and Syrdal and Gopal (1986) that include some transform of the distance between $F_1$ and $f_0$ as a determinant of vowel identity.

An issue that goes unaddressed in a fair amount of the work in this area is the very basic question of why $f_0$ should play any role at all in vowel identification. In acoustic terms, distinctions between one vowel and another are determined largely by the filter, so it is not immediately obvious why $f_0$ should play any role in the perception of vowel identity. The most common (but not sole) explanation for the influence of $f_0$ on vowel identification
suggests that it is a psychological effect in which decision criteria are adjusted based on learned associations between $f_0$ and formants; i.e., higher $f_0$s lead listeners to expect higher formant frequencies and vice versa (e.g., Assmann and Nearey, 2008; but see Irino and Patterson, 2002; Patterson et al., 2008, for a very different explanation). Scale factors that relate different talker groups are substantially larger for $f_0$ than they are for formants. Using the PB data, average scale factors relating men and women are 1.16–11.19 for the first three formants, but nearly an octave (1.71) for $f_0$. Nevertheless, average $f_0$ values are strongly correlated with average values for $F_1$, $F_2$, and $F_3$ ($r = 0.82–80.87$).

There is a good deal of evidence showing that vowels are less intelligible when typical relationships between $f_0$ and formants are altered. For example, Slawson (1968) showed that changes in the quality of synthetic vowels occur when $f_0$ is shifted by an octave while leaving the formants unchanged. However, the number of shifts in vowel identity was significantly reduced when an increase in formant frequencies of just 10% accompanied the one-octave $f_0$ shift.

The most complete set of findings on this problem come from a series of studies by Assmann and colleagues (Assmann and Nearey, 2007, 2008; Glidden and Assmann, 2004). For example, Assmann and Nearey (2008) used a high-quality source-filter synthesizer (Kawahara, 1997; Kawahara et al., 1999) to resynthesize vowels excised from /hVd/ syllables spoken by men, women, and children. The spectral envelope and $f_0$ were then manipulated alone or in combination. Identification accuracy declined with increases in either envelope or $f_0$; however, identification rates remained high when both $f_0$ and envelope were increased in a manner that is consistent with the observed correlation between $f_0$ and formants. As Assmann and Nearey note, “performance is poorest in conditions where the formant pattern and $f_0$ violate expected patterns in natural speech” (p. 3205). A second experiment using utterances resynthesized from men, women, and children produced generally similar findings. While men’s vowels were more vulnerable to downward frequency shifts and children’s vowels were more vulnerable to upward shifts, the basic finding remained the same as the experiment using only adult male talkers: Labeling performance declined as the frequency-shifted utterances moved further away from a regression line relating $f_0$ and formant frequency.

The experiments by Assmann and colleagues, along with a series of related experiments by Smith, Patterson, and colleagues (e.g., Smith et al., 2005, 2005; Smith et al., 2007; Patterson et al., 2008), which are not discussed here, are consistent with the idea that $f_0$ plays a role in intrinsic normalization. However, there remain aspects of this idea that are not at all straightforward. As noted earlier, the most common explanation for the $f_0$ effect is based on learned associations across talkers between average $f_0$ and average formants (or some as-yet unspecified measure of “envelope height” based on the shape of the envelope). However, in recognizing speech in real-time based on intrinsic factors, listeners must make use of $f_0$ and formants over some short time interval rather than average values of these parameters. The consequences of this simple observation are not trivial. For example, even when restricting our attention to citation-form syllables, correlations between $f_0$ and formants are not very strong; e.g., using the PB data, correlations between $f_0$ and formants calculated across vowel type are just 0.22, 0.28, and 0.57 for $F_1-F_3$, respectively. Further, ordinary conversational speech spoken by a single talker shows variation in $f_0$ extending over as much as an octave (Lieberman, 1967). Nearey (1989) asked whether it is possible that talkers adjust $F_3$ as $f_0$ varies in order to maintain a more nearly constant $F_1-f_0$ distance. The evidence for this is not particularly strong (Syrdal and Steele, 1985). (For an unusually thoughtful discussion of this and other issues related to the possible role of both $f_0$ and $F_3$ in vowel identification, see Nearey, 1989).
Extrinsic normalization

The basic idea underlying extrinsic normalization is that vowels are identified not simply by determining the position of that vowel in acoustic-phonetic space but rather by determining the relationship between a to-be-recognized vowel and a frame of reference that is established by listening to other vowels spoken by that talker. This idea was proposed by Joos (1948), who noted:

On first meeting a person, the listener hears a few vowel phones, and on the basis of this small but apparently sufficient evidence he swiftly constructs a fairly complete vowel pattern to serve as a background (coordinate system) upon which he correctly locates new phones as fast as he hears them.

(Joos, 1948, p. 61)

The idea is illustrated in Figure 9.16 (modeled after Barreda, 2013). The left panel shows formant frequencies for a typical adult male (based on PB averages) while the right panel shows the same kind of data for a typical child. The dot in each panel is an unknown vowel with $F_1 = 580$ Hz and $F_2 = 1220$ Hz. It can be seen that the unknown vowel is quite close to /ʌ/ in relation to the man’s vowels but close to /ʊ/ for the child. Evidence that is relevant to extrinsic normalization comes from both modeling studies that test the feasibility of various algorithms for identifying vowels using extrinsic normalization and, more directly, from listening studies that seek to determine whether listeners appear to behave as though they are identifying vowels in relation to a frame of reference that is established by listening to other utterances spoken by a given talker.

A great deal of modeling work has been directed at developing extrinsic normalization methods. A small but reasonably representative sampling of this work will be described here. A widely cited method proposed by Gerstman (1968) for use in automatic speech recognition uses a simple linear rescaling of $F_1$ and $F_2$ values based on the minimum and

---

**Figure 9.16** Average log-transformed formant frequency values for men (left) and children (right) based on the data from Peterson and Barney (1952). $F_1$ values vary between 270 and 1130 Hz and $F_2$ values vary between 845 and 3200 Hz. The dot in each figure represents a vowel of unknown identity with $F_1$ at 580 Hz and $F_2$ at 1220 Hz. If the vowel is recognized based on its relationship to each of the reference frames, it might be heard as /ʊ/ based on the reference frame on the left but /ʌ/ based on the reference frame on the right.

*Source:* The figure was modeled after Barreda (2013).
maximum $F_1$ and $F_2$ values for each talker, independent of the vowel type that was spoken. Using the PB data – from unanimously identified vowels only – Gerstman reported that 97.5% of the vowels were classified correctly. This percent correct figure is slightly better than the 94.5% intelligibility shown by PB’s listeners, although the algorithm was presented with measurements from unanimously identified words only, meaning that, by definition, the appropriate comparison figure for listener is 100%. Nevertheless, performance of the Gerstman algorithm was quite good.

A method developed by Lobanov (1971) is quite similar to Gerstman, except that formant frequencies are converted to $z$ scores rather than proportions. The method was tested using Russian vowels spoken by three men in a wide variety of phonetic contexts. The method was successful in both minimizing within-vowel-category variability and maximizing the distances between vowel categories. Although the number of talkers used in the test was surprisingly small given that the main purpose was to address acoustic differences across talkers, the Lobanov procedure has been used successfully in a variety of areas of phonetics (e.g., Adank et al., 2004). A closely related method is the log-mean procedure described by Nearey (1978), which is also in common use.

Many other extrinsic models have been developed over the years, but as Nearey (1992) and others have noted, these data-analytic approaches to the problem can suggest logically possible perceptual strategies, but the findings can tell us little about whether listeners actually use strategies such as these. Further, some of these algorithms were explicitly developed to address problems in areas such as automatic speech recognition and were never intended to model human perception (e.g., Gerstman, 1968). In that context, listening studies have a more direct bearing on the role that might be played by extrinsic normalization. The best known study in this area is a demonstration by Ladefoged and Broadbent (1957) of what certainly appears to be extrinsic normalization of the kind that is illustrated in Figure 9.16. The authors synthesized four test words (“bit,” “bat,” “bet,” “but”) which were preceded by six versions of an introductory carrier phrase (“Please say what this word is . . .”). The six versions, which were intended to simulate different talkers, consisted of a standard phrase and five other versions with formant frequencies shifted in various ways ($F_1$ down, $F_1$ up, $F_2$ down, $F_2$ up, $F_1$ down, and $F_2$ up). Listeners’ identification of the test words was strongly affected by the acoustic characteristics of the carrier phrase, consistent with the view that, “the phonetic quality of a vowel depends on the relationship between the formant frequencies for that vowel and the formant frequencies of other vowels pronounced by that speaker” (Ladefoged and Broadbent, 1957, p. 99). A brief demonstration of the Ladefoged and Broadbent experiment can be found at this link: https://engineering.purdue.edu/~malcolm/interval/1997-056/VowelQuality.html

There is a good deal of other experimental work that is consistent with the idea that listeners make use of information about a talker’s speech in recognizing subsequent utterances spoken by that talker. For example, Creelman (1957) found that recognition accuracy for words presented in noise was significantly lower when the identity of the talker was unpredictable from one trial to the next. Similarly, a word recognition reaction time study by Summerfield and Haggard (1975) found longer latencies in word lists with mixed versus single talkers. The increased latencies are thought to reflect the processing overhead associated with talker normalization. In one of their conditions, Kato and Kakehi (1988) asked listeners to identify a series of Japanese monosyllables spoken by different talkers and presented in noise with an 8 dB signal-to-noise ratio. Recognition accuracy increased linearly from 70 to 76% from the first to the fifth syllable, with no further increases beyond that. The effect was not especially large in absolute terms, but it was statistically reliable. Finally, a series of experiments
by Mullennix et al. (1989) showed that word recognition accuracy decreased and response latencies increased when talkers were mixed (seven men, eight women) versus segregated. Further, they reported significant effects for mixed vs. segregated talkers when the words were presented in noise (S/N = +10, 0, and −10 dB) and when they were presented in quiet.

An unusually interesting paper by Eklund and Traunmüller (1997) found that identification error rates for whispered Swedish vowels spoken by men and women were nearly five times higher on trials in which the sex of the speaker was misidentified than on the remaining trials. This would seem to be a clear indication that their listeners were using a decision about the sex of the talker to identify the vowel. This finding is consistent with the idea that vowel identity is affected by listener judgments about the characteristics of the talker, a specific form of extrinsic normalization. However, Eklund and Traunmüller found that the reverse was also true: The error rate for identifying speaker sex was four times higher on trials in which the vowel was incorrectly identified. In phonated speech, the two most important cues in distinguishing the voices of men and women are \( f_0 \) and formants (e.g., Assmann et al., 2006; Hillenbrand and Clark, 2009), but in whispered vowels the only major cue to speaker sex is the spectrum envelope (or formants derived from it). The conclusion that appears inescapable is that vowel identification depends on a decision about speaker sex and that identification of the sex of the speaker depends on a decision about vowel identity. It is not obvious how this circular dependency might be implemented either by neurons or by a computer algorithm.

A frequently cited paper by Verbrugge et al. (1976) produced mixed results concerning extrinsic normalization. In one experiment, listeners identified /hVd/ syllables with 15 vowel types (the vowel types from H95 plus /ː/, /œ/, and /ɐ/) spoken by 30 talkers (men, women, and children). In one condition, the talkers were presented in random order with no precursor, while in a point-vowel precursor condition the test syllables were preceded by /kip/, /kup/, and /kup/ spoken by the same talker as the test syllable. (The idea that the point vowels may play a special role in talker normalization goes back at least as far as Joos, 1948.) The error rate for the no-precursor condition (12.9%) was not reliably different from that of the precursor condition (12.2%). In a follow-up experiment, listeners were asked to identify /pVp/ syllables under four conditions: (a) Isolated syllables, mixed talkers; (b) isolated syllables, utterances grouped by talker; (c) test utterances from the mixed-talker set preceded by /hV/ syllables with the point vowels (/hɪ/, /hæ/, /hʌ/); and (d) test utterances preceded by /hV/ syllables with non-point vowels (/hɪ/, /hæ/, /hʌ/). Results were mixed. The error rate was significantly higher for the mixed-talker condition (17.0%) than the segregated-talker condition (9.5%). However, exposure to a talker’s speech did not result in a significant decrease in error rate for either point-vowel precursors (15.2%) or non-point-vowel precursors (14.9%). (See Morton et al., 2015, for different results on the effectiveness of precursors.) Further, there was no reliable difference between point-vowel and non–point-vowel precursors. In a third experiment, listeners were asked to identify /pVp/ syllables that had been excised from a rapidly spoken sentence. A relatively high error rate of 23.8% was reported, but the error rate for a separate group of listeners tested on syllables that were preceded by same-talkers point-vowel precursors was significantly higher (28.6%). The authors suggested that the citation-form precursors led listeners to expect a more leisurely speaking rate which clashed with the briskly spoken /pVp/ syllables that had been excised from a sentence. The lowest error rate (17.3%) was observed in a third condition in which the test syllables were presented in the original sentence context. The authors concluded that the sentence frame, “aids vowel identification by allowing adjustment primarily to a talker’s tempo, rather than to the talker’s vocal tract” (Verbrugge et al., 1976, p. 198).
In a set of natural speech control conditions for an experiment on the perception of sin-
ewave vowels, Hillenbrand et al. (2008) found no effect at all for an introductory carrier
phrase (CP) on the identification of 16 monophthongs and diphthongs spoken by 20 talkers
(10 men, 10 women). Listeners identified /hVd/ test syllables presented under three condi-
tions: (a) Isolated syllables, (b) a within-talker carrier phrase (CP) condition in which test
syllables were preceded by a CP (“The next word on the list is . . .”) spoken by the same
talker who uttered the test syllable, and (c) a cross-talker condition in which the talkers who
produced the CP and test syllables were paired randomly (with 20 talkers, this means that
the CP and test-word talkers would differ on 95% of the trials). Identification rates for the
three conditions were uniformly high and statistically indistinguishable from one another
(isolated syllables: 95.5%, within-talker CP: 95.3%, across-talker CP: 96.4%).

A set of results that are only rarely cited within the context of extrinsic normalization is
the commonplace finding that vowels spoken by a diverse group of talkers, and presented to
listeners in random order, are highly identifiable. For example, consider the 94.5% identifi-
cation rate for 10 vowel types spoken by 76 talkers in PB, the 95.4% identification rate for
12 vowel types spoken by 139 talkers in H95, and the ~96% identification rate reported by
Abramson and Cooper (1959; cited in Verbrugge et al., 1976). Verbrugge et al., are among
the few to comment on this point:

there is reason to doubt whether a preliminary normalization step plays the major role
in vowel perception that is commonly attributed to it. Remarkably low error rates have
been found when human listeners identify single syllables produced by human talkers.
Peterson and Barney (1952) and Abramson and Cooper (1959) found average error
rates of 4% to 6% when listeners identified the vowels in h-vowel-d words spoken in
random order by a group of talkers. The test words were spoken as isolated syllables
and in most conditions the listeners had little or no prior experience with the talker’s
voice. On the face of it, these low observed error rates seem inconsistent with any
theory that stresses the need for extended prior experience with a talker’s vowel space.
(Verbrugge et al., 1976, p. 199)

It is not a particularly simple matter to determine what the literature described in this sec-
tion has to tell us about the status of extrinsic normalization. On the one hand, it appears
undeniable that listeners are both more accurate and faster in recognizing speech when
talkers are segregated rather than mixed. These effects tend not to be especially large, but
they are real enough. On the other hand, it is not entirely clear what information listeners
derive from the speech of these talkers, or how that information is put to use in recognizing
subsequent speech. If utterances from a talker are used to derive a reference frame that is
used in the manner described by Joos (1948), Ladefoged and Broadbent (1957), and others
(and illustrated in Figure 9.16), then how could Verbrugge et al. (1976) – in two separate
experiments – have failed to find that same-talker precursors improve vowel identification
performance? Similarly, what explains why Hillenbrand et al. (2008) did not find even a hint
of a difference in vowel recognition when comparing isolated syllables, syllables preceded
by a within-talker carrier phrase, and syllables preceded by a cross-talker carrier phrase?
Finally, as noted by Verbrugge et al., if the establishment of a reference frame is needed
in order to recognize speech from a variety of talkers, what explains the very high vowel
identification rates of ~94–96% when words spoken by large groups of men, women, and
children are presented in random order, as in PB and H95?
These negative findings with natural-speech carrier phrases or precursors might be seen as an indication that listeners rely on extrinsic normalization only when the speech has been degraded in some way, such as noise, reverberation, or filtering. However, Mullennix et al. (1989) reported clear evidence for increased accuracy and shorter latencies for segregated vs. mixed talkers when test words were presented in noise and when they were presented in quiet. More important, the central idea underlying extrinsic normalization of the kind proposed by Joos (1948) and Ladefoged and Broadbent (1957) is that a reference frame is needed to resolve the kind of ambiguity that is illustrated in Figure 9.16. Ambiguity problems such as these do not vanish under good listening conditions, so it is not easy to understand why listeners would fail to make use of an available reference frame when identifying clearly spoken utterances. There are, of course, the Ladefoged and Broadbent findings, which appear to show exactly these kinds of effects, but it is not at all obvious why these kinds of results should be seen only with synthetically generated speech and/or when talker variation is simulated by introducing somewhat arbitrary shifts in formant frequencies. A solution to these problems will not be offered here, but resolving these kinds of questions would appear to be a fit topic for further work on talker normalization.

Chapter summary

This chapter begins with an overview of source-filter theory as it applies to the production of phonated vowels. This is followed by a description of the spectral characteristics of NAE vowels, and the identification of these vowels by listeners. The most striking results, seen in studies such as PB and H95, are two findings that appear to be at odds with one another: (a) Acoustic measurements show a good deal of variability across talkers, and significant overlap among adjacent vowel categories; but (b) when presented to listeners for identification in random order, the utterances prove to be highly intelligible, with error rates of ~4.5–5.5%. These findings indicate that the most common method that is used to represent the acoustic properties of vowels – the two lowest formant frequencies – is either wrong or incomplete, and/or that there is more to the perceptual mechanisms underlying vowel identification than was initially imagined. The remainder of the chapter focuses on four research areas that have attracted a good deal of attention in the literature on vowel recognition: (a) Formants versus spectral shape, (b) the role of spectral change, (c) the role of duration, and (d) talker normalization.

Formants vs. spectral shape

Formant theory, which is more often implicitly assumed rather than explicitly stated, suggests that vowel identity is controlled by the frequencies of the two or three lowest formants rather than the detailed shape of the spectrum. Spectral shape or whole spectrum models, on the other hand, assume that vowel identity is controlled by the overall shape of the spectrum envelope. As one example of a spectral shape model, Bladon and Lindblom (1981) used an auditory model to derive critical band spectra for a set of synthetic vowels. An auditory distance metric was then used to predict perceptual distances between pairs of critical band spectra. The perceptual distances derived from the model correlated strongly with perceived distances between the same pairs of vowels. However, in what we believe is the most important experimental work on this question, Klatt (1982) showed conclusively, in three separate experiments, that the only way to induce a large change in the phonetic quality of synthetic vowels was to alter the formant-frequency pattern. Other changes in
spectral shape, while quite audible to listeners, have very little effect on judgments of the phonetic distance between pairs of vowels. Klatt suggested that the evidence argued in favor of a spectral shape model that was dominated by high-energy regions of the spectrum, but without requiring formant tracking. Klatt proposed a phonetic distance metric that compared spectral slope rather than amplitude differences between input and reference spectra, with the goal of representing similarities and differences in peak frequencies rather than amplitude. Klatt reported a strong correlation \(r = 0.93\) between his weighted slope measure and listener judgments of phonetic distance from the first of this three listening experiments.

Consistent with Klatt’s findings are vowel classification findings using a recognition algorithm that compared narrow band input spectra with a set of smoothed spectral shape templates that were empirically derived by summing the narrow band spectra of like vowels spoken at similar time points throughout the course of the vowel (Hillenbrand and Houde, 2003). In preparing narrow band spectra for template generation, and to serve as input to the classifier, steps were taken to emphasize spectral peaks at the expense of spectral shape details such as formant amplitudes, overall spectral tilt, and spectral details in between harmonics. Using separate templates for men, women, and children, the algorithm classified 1,668 vowels from H95 with 91.4% accuracy. Significantly, the performance of the algorithm fell dramatically to just 59.9% when a simulation of masking was removed from the classifier. The masking simulation was designed to emphasize spectral peaks at the expense other aspects of spectral shape which contribute minimally to judgments of phonetic quality. Overall, findings in this area do not argue in favor of either formants or spectral shape, as the concept is typically conceived (e.g., Bladon and Lindblom, 1981). Although opinions on this issue vary, the argument made here is that the evidence is consistent with the idea that listeners base their judgments of vowel identity on a representation of spectral shape that is dominated by energy at or near spectral peaks, but without requiring the assignment of spectral peaks to specific formant slots \(F_1, F_2, \ldots\) – i.e., without tracking formants.

The role of spectral change

Researchers were aware of the potential importance of spectral change in vowel identification as early as the 1950s (e.g., Potter and Steinberg, 1950; Peterson and Barney, 1952; Tiffany, 1953). For the NAE vowels that have received the most attention, this idea now has the status of established fact. Evidence comes from four overlapping sources. First, individual NAE vowel types have distinctive patterns of spectral change. A few vowel types, particularly /i/ and /u/, show little spectral movement throughout the course of the vowel, but most others show diphthong-like movements called vowel inherent spectral change. Evidence from several statistical pattern classification studies shows that vowels can be classified with considerably greater accuracy when spectral change is included among the classification features (e.g., Zahorian and Jagharghi, 1993; Hillenbrand et al., 1995, 2001). Second, a series of experiments using silent-center vowels showed vowel identification accuracy was unaffected when vowel centers were excised from CVC syllables (e.g., Jenkins et al., 1983; Parker and Diehl, 1984; Nearey and Assmann, 1986; Andruski and Nearey, 1992; Strange et al., 1994). These experiments indicate that the stationary targets that might have been thought of as vowel identification are, in fact, not necessary. Third, several experiments indicate that static targets are not merely unnecessary, they are also insufficient for vowel identification. The most striking piece of evidence on this point comes from a remarkable study by Fairbanks and Grubb (1961), described at considerable length earlier. Briefly, the authors went to great lengths to record the highest quality examples of eight
static vowel types spoken by just seven talkers, all men and all faculty members in Speech and Hearing Science. Elaborate procedures were in place to audition and re-record any utterances that were judged to be unsatisfactory. In spite of these and other steps that were taken to obtain the highest quality recordings, the average intelligibility was just 74%, much lower than the figures reported in PB and H95, who recorded more vowel types, with a much greater diversity of talkers, and without the elaborate auditioning procedures used by Fairbanks and Grubb. The main explanation for the high error rate in Fairbanks and Grubb is that the signals did not show the natural VISC patterns that are typical of NAE speech, although the use of fixed-duration vowels almost certainly played some role as well. Other evidence that static targets are not sufficient to explain the excellent intelligibility observed in studies such as PB includes Hillenbrand and Gayvert (1993), who used a formant synthesizer to generate 300-ms static versions of all 1,520 signals from PB based on their measurements of $f_0$ and $F_1$–$F_3$. The average intelligibility was 73.8%, similar to the figure reported by Fairbanks and Grubb. Similarly, Hillenbrand and Nearey (1999) asked listeners to identify three versions 300 /hVd/ drawn from the H95 signals: (a) Naturally spoken signals (NAT), (b) formant-synthesized versions generated from the original formant contours (OF), and (c) formant-synthesized versions generated from flattened formants (FF). The key OF–FF comparison showed a very large drop in intelligibility from 88.5% for the OF condition to 73.8% for the FF condition. In summary, the body of evidence in this area is quite clear in showing that NAE vowels are much better thought of as distinctive trajectories through acoustic-phonetic space rather than points in that space.

The role of duration in vowel identification

Although vowel duration is not phonemic in English, there is good evidence that listeners attend to duration in identifying vowels. Early studies using isolated, sustained vowels showed that listeners make use of vowel duration in a manner that is generally consistent with measured differences in inherent duration. For example, Tiffany (1953) showed that long vowels such as /ɑ/ and /e/ were more likely to be identified as the intended vowel when presented at longer durations, and vice versa for vowels with short inherent durations. Similar findings were reported by Stevens (1959), with the notable exception that the effect of duration on /i/–/ɪ/ and /u/–/ʊ/ was relatively weak, in spite of large and reliable inherent duration differences for these pairs. A more recent study (Hillenbrand et al., 2000) used /hVd/ syllables showing natural VISC patterns that were synthesized at original, shortened, and lengthened vowel durations. The original-duration signals were highly intelligible (96.0%); shortening or lengthening the vowels resulted in nearly symmetrical decreases in intelligibility of about five percentage points. However, this relatively modest drop in intelligibility was arrived at by averaging some cases in which vowels produced rather large effects (e.g., /æ/–/ɛ/ and /ɑ/–/ɔ/–/ʌ/) with other cases – notably /i/–/ɪ/ and /u/–/ʊ/ – in which duration produced almost no effect at all, in spite of large differences in inherent duration. Pattern recognition findings suggested that a likely explanation is that listeners give little weight to duration for vowels that can be identified well based on spectral features alone, and vice versa.

Talker normalization

The acoustic characteristics of a given vowel type vary across talkers, even when dialect is held constant. To many investigators, the fact that listeners have little difficulty recognizing
the speech of many talkers suggests the operation of talker normalization mechanisms. Following Ainsworth, we distinguish between intrinsic normalization, in which all normalizing information is available directly in the to-be-recognized utterance, and extrinsic nomination, in which previous speech from a talker is used to establish a reference frame that aids in the identification of later utterances.

_Intrinsic normalization_

As Miller (1989) noted, many intrinsic normalization schemes are variations of Lloyd’s (1890, 1892) relative resonance theory, which suggests that vowels with similar qualities have similar formant ratios. Influential models developed by Miller (1984, 1989) and Syrdal (1985; Syrdal and Gopal, 1986) added a transform of the fundamental frequency to formant ratios to disambiguate vowel pairs with similar ratios. In addition to modeling work of this kind, there is clear evidence from listening experiments showing that judgments of vowel identity are affected by variations in $f_0$ when formants are held constant (e.g., Miller, 1953; Slawson, 1968; Fujisaki and Kawashima, 1968; Ryalls and Lieberman, 1982). There is disagreement about the critical issue of why $f_0$ should affect vowel recognition. A common (but note sole) suggestion is that $f_0$ plays an indirect role in vowel identification by affecting decision criteria based on learned associations between $f_0$ and formant frequencies (e.g., Assmann and Nearey, 2008; Barreda, 2013). These findings are consistent with the idea that $f_0$ plays an indirect role in vowel identification that is based on correlations between $f_0$ and formants in natural speech. However, as noted earlier, the idea is not without problems, including: (a) Even in citation-form speech, the relationship between $f_0$ and formants at the level of individual tokens (as opposed to averages across talkers) is not especially strong; and (b) instantaneous $f_0$ varies quite substantially in ordinary conversational speech.

_Extrinsic normalization_

One of the earliest pieces of experimental work in this area turns out to be one of the more interesting. Using a synthesizer that is far from the current state-of-the-art, Ladefoged and Broadbent (1957) created four test words that were preceded by either a standard introductory carrier phrase or one of five formant-shifted versions representing different talkers. Identification of the test words was affected by the acoustic properties of the carrier phrase in a manner that was generally consistent with the reference-frame hypothesis proposed by Joos (1948). Another major piece of evidence favoring a role for extrinsic normalization is the finding, replicated in many studies, that word recognition accuracy is higher, and/or latencies lower, when talkers are segregated versus mixed (e.g., Creelman, 1957; Summerfield and Haggard, 1975; Kato and Kakuei, 1988; Mullenix et al., 1989). Evidence from other studies has been either mixed or negative. For example, Verbrugge et al. (1976) replicated the common finding of higher recognition accuracy for segregated than mixed talkers, but found no evidence for higher recognition rates when test utterances were preceded by same-talker precursors consisting of either point- or non-point vowels. If the segregated-talker advantage is explained on the basis of extrinsic normalization, what accounts for the failure to find an improvement for same-talker precursors? In discussing these issues, Verbrugge et al., also asked the straightforward question of how extrinsic normalization could be essential to vowel recognition in light of the very high recognition rates of ~95% from studies such as PB in which utterances from a diverse group of talkers are presented for identification in random order. Finally, using naturally spoken utterances, Hillenbrand et al.
(2008) found no difference in vowel recognition rates when carrier phrases were spoken by
the same talker as the test syllable versus a different talker. Some of these findings would
appear to be in conflict with others. Unfortunately, we are not in a position to offer a solution
to these apparently conflicting findings.

Future directions
Vowel perception research has been dominated by studies using citation-form utterances
such as isolated vowels or syllables/words with a simple and often fixed structure. Further,
these utterances are typically longer and are almost certainly spoken with less pronounced
coarticulatory patterns than are seen in ordinary conversational speech. There is an obvious
reason for these constraints on speech material: These simpler utterances allow attention to
be focused on a small and clearly specified set of underlying physical parameters. However,
there are certain problems that cannot be understood with these kinds of utterances. One
example, described earlier, concerns the influence of duration on vowel identity. The results
of this work using isolated vowels and syllables seem clear enough, but in connected speech
there is a large and complex set of factors that exert a systematic influence on segment
duration (e.g., Klatt, 1976). It is not at all clear how listeners go about assigning duration
variability to specific factors such as vowel identity, final-consonant voicing, speaking rate,
word- and sentence-level stress, and the many other factors that influence vowel duration.
A similar problem concerns how listeners go about assigning spectral change patterns to
either vowel identity (i.e., treating the spectral changes as VISC patterns) or coarticulation.
Research along these lines using connected speech would clearly require some ingenuity,
but it would represent a useful avenue of research.

Notes
1 We are not suggesting that the signal processing steps described here literally take place when lis-
teners recognize speech. These steps are intended as a proxy for psychological rather than physical
processes that are involved in perception. For example, Klatt’s (1982) findings clearly show that
features such as spectral and formant amplitude relationships are quite audible to listeners, so it
cannot be that these features are literally removed from the input signal. However, these kinds of
features play a limited role in judgments of phonetic quality.
2 We experimented with methods for normalizing input spectra and templates across talkers (Houde
and Hillenbrand, 2007). These methods worked reasonably well, but we have not yet found any-
thing that performs as well as separate templates for men, women, and children.
3 As will be discussed later, there is clear evidence that listeners make use of vowel duration in iden-
tifying vowels, indicating that at least some of the performance advantage for listeners is related to
the fact that our algorithm does not incorporate duration information. We have not yet developed
a method for combining the spectral information that is used in the current model with duration
values.
4 Nearey (1989) makes the point that it is possible to argue that intrinsic schemes do not constitute
normalization at all: “Although it is possible to formulate such approaches as normalization pro-
cedures the term is rarely used by this group. Instead, the invariance problem is deemed not to
exist when the correct parametric representation of spectral properties of vowels is considered.”
(p. 2080).
5 Irino and Patterson (2002) have proposed that an obligatory transformation of input stimuli to a
sensation of size takes place at an early stage of auditory analysis, and that this size sensation is not
specific to speech or to human listeners. Further, Smith et al. (2005) have argued that the effects
of frequency shifts on vowel identity cannot be explained on the basis of learned associations between \( f_0 \) and formants. See Assmann and Nearey (2008) for a discussion of this question.

6 The authors made what would seem to be surprising choices in the approach that was adopted for creating the five variations on the standard talker. In the formant data from both PB and H95, it can be seen that the main (though not sole) source of variation across individual talkers lies on a lower-left-to-upper-right diagonal, indicating the kinds of near-the-regression-line signals that produced the best intelligibility in the scaling studies by Assmann and colleagues. However, the manipulations that produced all five simulated talkers were in directions that moved the formants away this regression line. It is not at all clear what effect, if any, this approach to simulating talker differences may have had on their findings. But given the impact that this very interesting paper has had in this area, it might be worth finding out.

7 Statistics for a condition-by-vowel interaction were not reported, but even a casual inspection of their data (see their table I) indicates that results were quite different from one vowel type to the next. While these differences were discussed, a compelling explanation was not offered.

8 In one sense, even these very high identification rates underestimate how intelligible these utterances are. In the PB data, fully 55% of the errors come from just two vowels – /ɑ/ and /ɔ/ – which are confused mainly either with one another, or with /ʌ/. If data from these two vowels are removed, the identification rate rises to 96.9%. Data from H95 are similar.

References


