7 Frequency selectivity and the perception of speech

S. ROSEN and A. FOURCIN

University College London, England

1 INTRODUCTION AND ORIENTATION

Of all the sounds experienced by human listeners, those of speech are certainly the most important. This primacy results from the central role which language plays in our lives, in its written as well as spoken forms. Indeed, most attempts to define what makes us specifically human invoke the use of language as the single most important criterion. Perhaps in some bygone, more dangerous age, with predators lurking unseen, were environmental sounds of warning more essential for survival. Today, most of us dwell in a world where car horns have become the major auditory warning, but modern communication systems, like the telephone, make ever-increasing use of our specific abilities to process spoken language. Current developments in speech synthesis, with prospects (some already achieved) of speaking clocks, calculators, cookers and washing machines, are likely to accelerate this trend.

This connection between speech and hearing does not, however, only flow one way. Although speech, as the most basic manifestation of language, must be seen as the finest expression of auditory function, it does not stand apart, uninfluenced by the fact that it is communicated via sound. In evolution, the senses of vibration and hearing preceded speaking, and so it is inescapable that basic auditory constraints moulded the way speech developed. In short, trying to understand the auditory processes involved in the perception of speech is likely to lead not only to a better understanding of hearing, but also, of speech itself.

On the other hand, it must not be supposed that purely auditory factors

Frequency Selectivity in Hearing
ISBN: 0-12-505625-7

will suffice completely to explain speech processing, which depends on a complex chain of events involving cognitive and productive processes, as well as sensory ones. Our aim in this chapter is simply to consider those aspects of speech perception that are illuminated when considered from the point of view of the hearing mechanism, with special reference to frequency selectivity. Previous chapters have shown the ubiquity and power of this basic concept in explaining many disparate phenomena, and we hope to build on those discussions. In addition, we attempt to explore the perceptual consequences of impaired frequency selectivity (so common in the hearing impaired, as discussed by Tyler in the previous chapter) and the implications for hearing-aid design. Apart from the purely scientific goal, it is likely that a greater understanding of the role of frequency selectivity in speech perception (both normal and impaired) will lead directly to improved auditory prostheses.

First, though, it will be necessary to describe in some detail the structure of these very special auditory stimuli—speech sounds.

2 BASIC ARTICULATORY AND ACOUSTIC PROPERTIES OF SPEECH

Unlike most of the stimuli used in auditory research, there is an intimate relationship between the acoustic properties of speech sounds, and their source—the human organs of speech. As the acoustic structure of speech is completely predictable (on physical principles) from the position and states of the organs of production, we begin by describing their structure and function.

A *Functional anatomy*

Figure 7.1 shows in broad outline the main structures which will concern us, in a frontal view. For convenience, we can consider the speech production system to be made up of three subsystems: the subglottal system, the larynx and the vocal tract.

The subglottal system, consisting of the trachea (or windpipe), bronchi, lungs and associated musculature, is by far the most important energy source for speech sounds. Air is pressed from the lungs and led via the bronchi and trachea to the larynx and vocal tract, where, for the majority of speech sounds, a constriction causes either quasi-periodic or random sound energy to be generated. Outside of this crucial function, the subglottal system is of little communicative importance. Although it can affect the acoustic structure of speech (in some second-order ways), there is no speech sound contrast that depends primarily on different subglottal gestures.

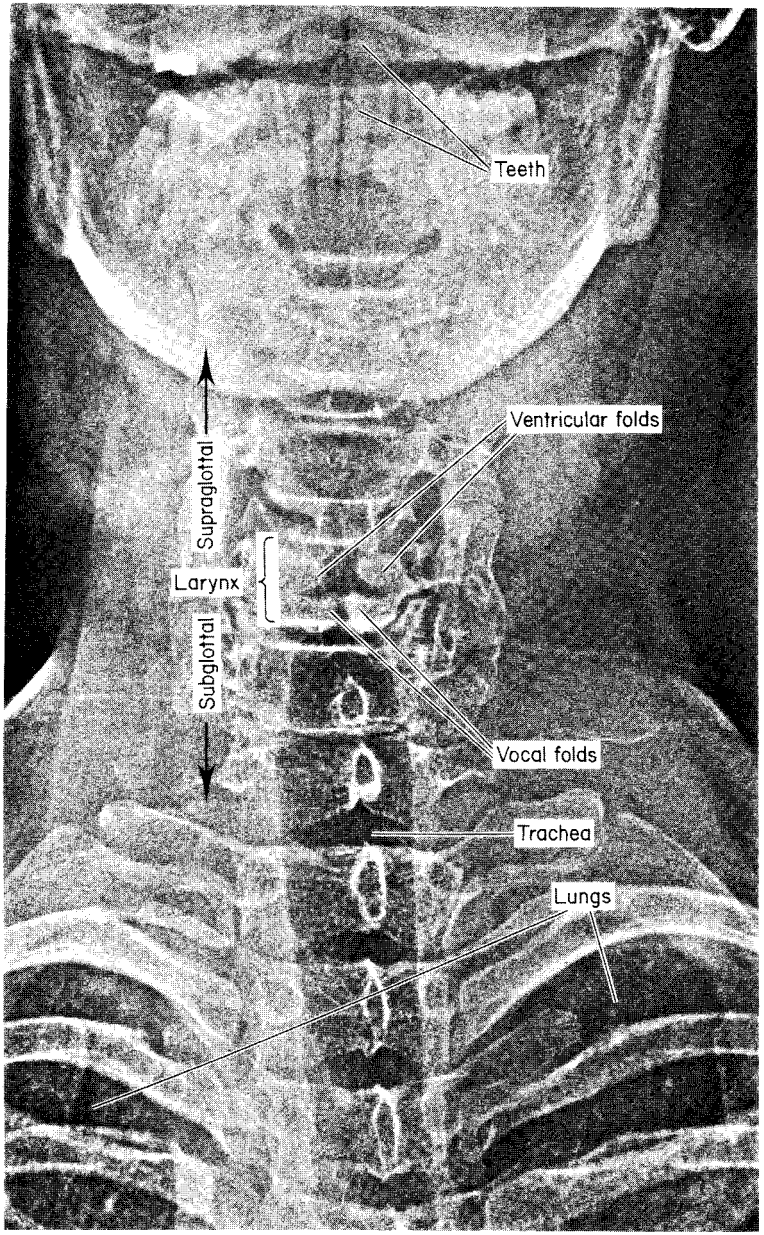


FIG. 7.1 The organs of speech production, as seen in a front-to-back xeroradiogram of an adult male. The vocal folds (in the larynx) are clearly seen. The space between them is called the glottis, and serves as the boundary between the sub- and supraglottal systems. The ventricular, or 'false' vocal folds (above the 'true' vocal folds), although prominent, play no important role in normal speech production. Not on this figure (but shown in Fig. 7.3) are the nasal cavities, which are an important feature of the supraglottal system. For further details, see the text.

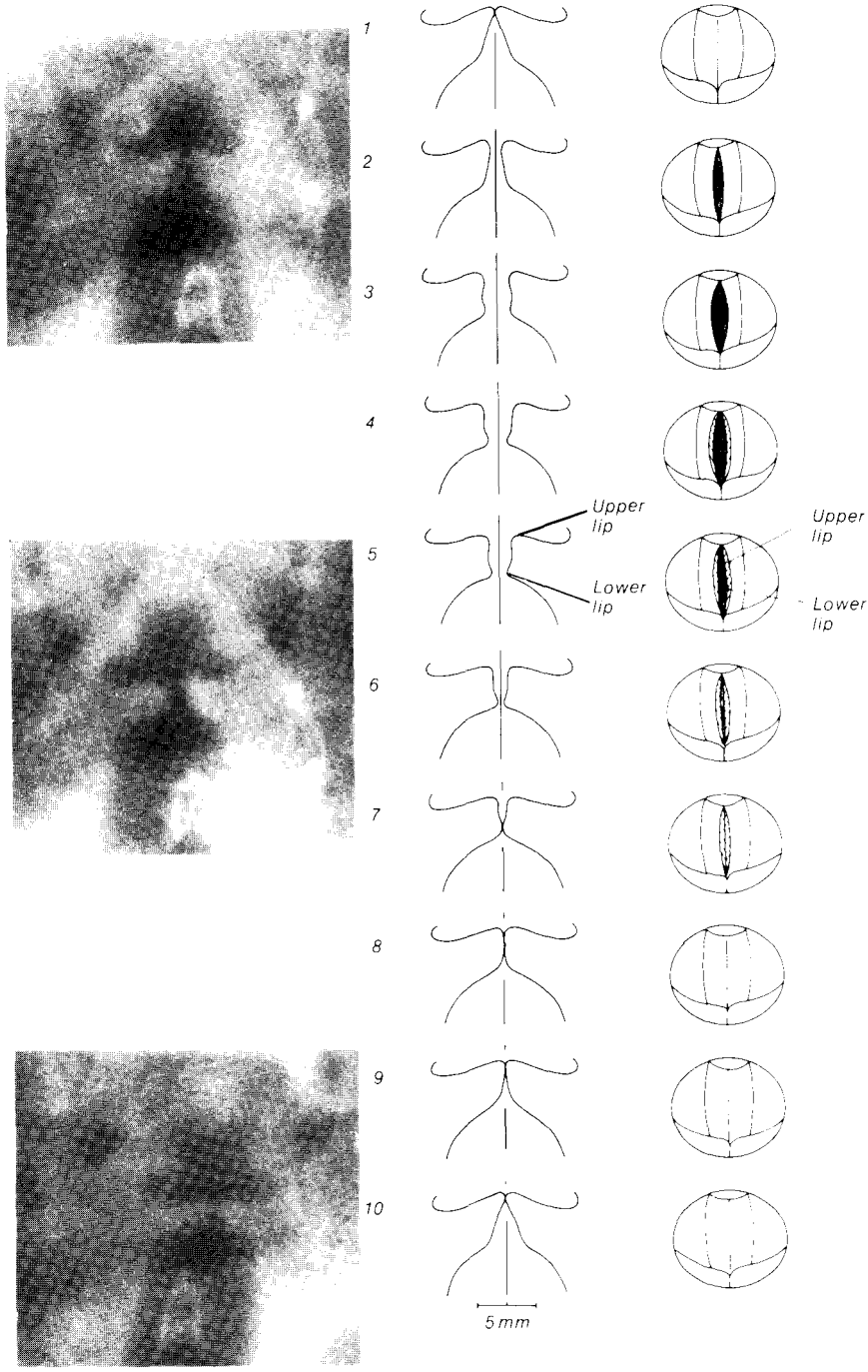
At the top of the trachea, we find the *larynx*, the primary site at which power transmitted from the lungs is transduced into sound. The actual generation of acoustic energy here depends on the *vocal folds* (or 'cords'), two elastic structures of muscle covered by a mucous membrane. These are stretched front-to-back on the lateral walls of the larynx, and can be controlled in position, tension and length. When the vocal folds are relaxed, and folded back for respiration, the space between them, (known as the *glottis*) is relatively large and allows a free flow of air. When they are adducted by the arytenoids and held sufficiently in tension so as not to vibrate, the glottal aperture thus formed can produce turbulence in the flow of air from the lungs. With the appropriate tension, however, they vibrate quasi-periodically when there is a sufficient flow of air. The sensation of *voice* is typically associated with this vibration.

The physical processes underlying vocal fold vibration are substantially explained by the aerodynamic-myoelastic theory of phonation (van den Berg, 1958, 1968; Titze and Talkin, 1979). Assume that the vocal folds are initially together. Air flowing from the lungs increases the subglottal pressure, until there is sufficient force to push the vocal folds apart. Now air can begin to flow through the glottis and into the vocal tract. As the air flow builds up, however, the air pressure in the glottis is reduced (the Bernoulli effect), which, along with the tension of the folds themselves, brings them together again, sharply. The cycle can then be repeated, with each closure producing a peak of acoustic energy (see Fig. 7.2). The voiced component of speech is associated with this larynx activity. It has the greatest intensity of all speech features and its presence characterizes human speech in all languages.

Since the folds normally vibrate quasi-periodically, the acoustic excitation to the vocal tract which they provide is also normally quasi-periodic. Thus, speech sounds that are voiced (that is, produced with the vocal folds vibrating) are normally heard as having a strong low pitch directly related to the frequency of this vibration. The vibration frequency is varied primarily through the tensioning of the vocal folds, but variations in their position and attitude, and in subglottal pressure, make secondary contributions.

Of course, we never hear the sound from the larynx directly. All sounds

FIG. 7.2 *(opposite)* The stages of vocal fold vibration, shown both schematically, and in X-ray flash photographs. The column furthest to the right shows a schematized view from above, while the middle column, a schematized view from the front. In the left column are three radiograms through the neck of an adult male, taken at times corresponding to the stages marked 2 (just after the opening of the folds), 6 (just as the folds are coming together) and 10 (during closure, just prior to the folds separating again). These radiograms were made with X-ray pulses of only 30-ns duration triggered by laryngograph electrodes (Fourcin, 1981) worn by the speaker. Note the white spots on the upper two radiograms caused by solder points on an electrode (Noscoe *et al.*, 1983). For further details, see the text. The two right-most columns are reproduced with permission and modifications from Hirano (1981).


1
2
3
4
5
6
7
8
9
10
Upper lip
Lower lip
Upper lip
Lower lip
5 mm

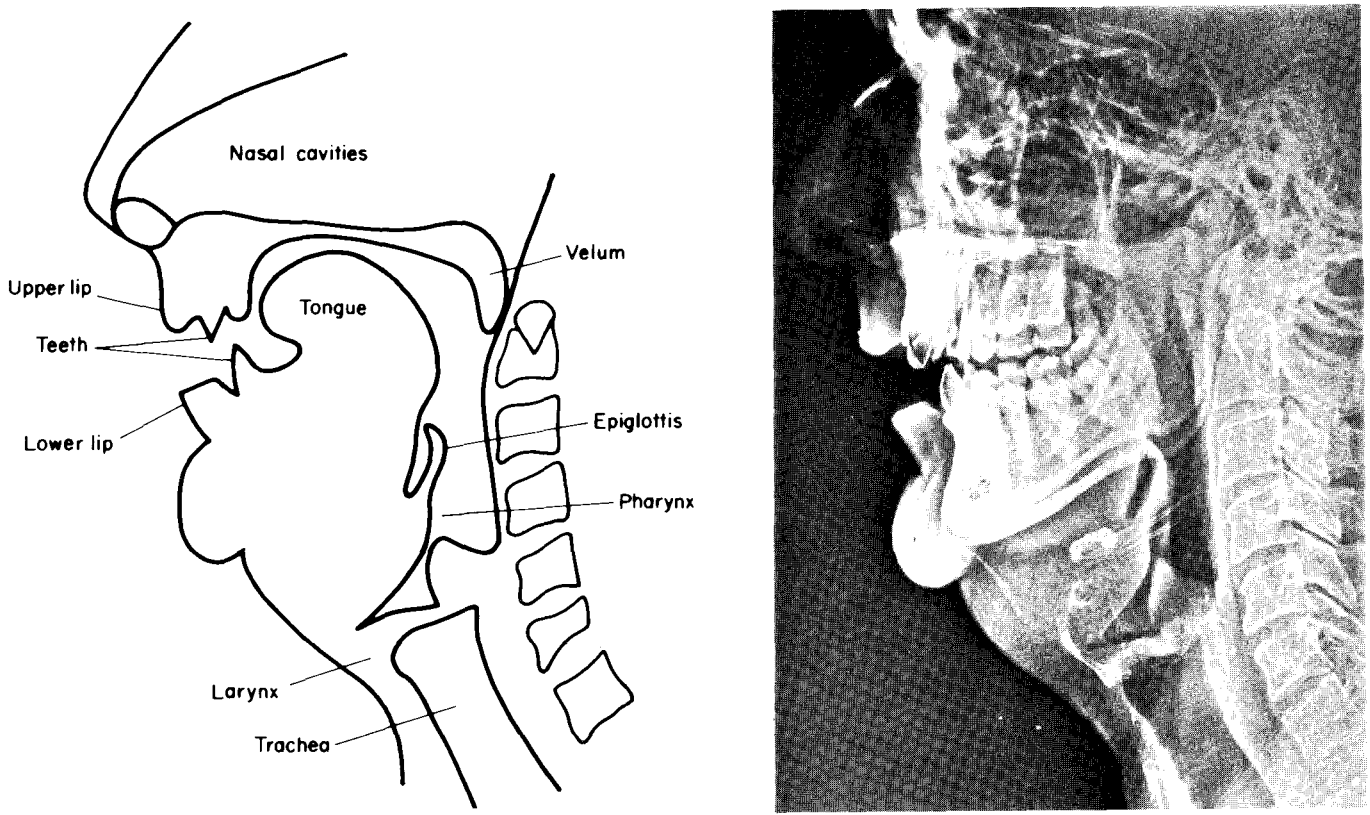


FIG. 7.3 The vocal tract. On the right is a xeroradiogram, from the side of an adult male uttering the vowel |i| 'ee'. On the left is a line drawing, traced from the xeroradiogram, in which the most important structures for speech production are labelled.

generated there are further modified by the cavities above the larynx—the vocal tract—seen in detail in Fig. 7.3.

If the vocal tract is relatively unconstricted, no new acoustic energy is generated along its length; it can only shape what is already available. The most straightforward case occurs in the production of steady-state non-nasalized vowels, when the soft palate, or *velum*, closes off the entrance to the nasal passages. The vocal tract can then be treated as a system of simple resonators which selectively shape the amplitude-frequency spectrum of the larynx output. The sounds produced by these alterations give percepts of various vowel qualities, with the response of the vocal tract (and hence the acoustic structure of the sound) being varied by changes in the positions of the relevant articulators (primarily the lips, jaw and tongue). Some continuant sounds (like English /l/ and /r/) can, from the acoustic point of view, be treated in a similar way, although the articulatory positions are rather more complicated than those found for simple vowels.

In many languages (e.g. Polish, French and Punjabi) there is at least one pair of vowels in which the lips, jaw and tongue are in similar positions, but which are distinguished by one vowel being nasalized and the other not. Nasalization occurs when the velum is lowered, opening, either partially or completely, the velar port connecting the nasal passages with the oral vocal tract. Then, the nasal cavities can exert their own influence, but in a different way from that of the oral tract. Although it is still a matter of the larynx output spectrum being shaped, now antiresonances can occur.

A similar effect occurs in the production of voiced nasal sounds (like English /m/, /n/ or /ŋ/ as in si*ng*). Although the velar port is open, there is a closure somewhere in the oral tract (in English at the lips, alveolar ridge or velum, respectively) which forces all the air flow through the nose. Again, there is a shaping of the spectrum of the larynx output through a set of resonances and antiresonances, but important antiresonances will be associated with the oral cavity formed between the point of vocal tract closure and the velar port.

So far we have only discussed sounds made with the oral and/or nasal tract relatively open, but there are many sounds for which a constriction is found in the vocal tract. Air flow through this constriction becomes turbulent leading to an aperiodic acoustic excitation which we hear as noise-like. Continuant sounds with this characteristic are known as *fricatives*. In voiceless fricatives, the aperiodic turbulence noise produced at the constriction is the sole acoustic energy source. The noise generated at the constriction is modified by resonances and antiresonances due to the cavities in front of and behind the constriction. The sizes of these cavities (and hence their acoustic properties and perceptual effects) are determined by the exact place in the vocal tract at which the constriction is made. Fricatives can be made with the constriction at a variety of places. In English, for example, the constriction for /s/ is

between the tongue and the alveolar ridge, while for /f/ it is between the lower lip and upper teeth.

Fricatives can also be voiced. The constriction is maintained, generating aperiodic energy, but the vocal folds are vibrating as well. Since the air flow from the lungs is modulated by the folds, so the frication noise is modulated at the rate of vocal fold vibration. This is known as *mixed excitation*.

Vowels, nasals and fricatives are all describable as sounds that are quasi-static, since they can be produced continuously. There are many speech sounds, however, which depend on dynamic changes for their identity. Simplest of these are the *diphthongs* (like b*ai*t or l*oud*) for which the articulatory position for one vowel moves continuously towards that appropriate for another. In other words, a diphthong is a vowel that changes with time. Similar in terms of production are the *glides* (like /w/ and /j/, as in *w*et and *y*et) save that the characteristic articulatory positions for glides are rather more constricted than those for diphthongs, and the transitions between states are faster.

Probably the most important inherently dynamic sounds are the *plosives* (e.g. /p/, /b/, /t/, /d/) for which a complex interplay between aerodynamic and mechanical forces occurs. All plosives are characterized by a complete closure somewhere in the vocal tract. For simplicity, let us consider in detail only English plosives followed by a vowel. After the closure (say at the lips as in /b/), air pressure builds up in the cavity behind, as the velum is normally closed too. When the plosive is released, a burst of air gives a transient excitation which will be spectrally modified by the vocal cavities. As far as is possible with a given constriction, the vocal tract will already be set for the following vowel before the release takes place; this 'co-articulatory' activity is basic to all aspects of speech, both productively and perceptively.

At release, the parts of the tract constrained by the constriction move quickly to the articulatory position for the following vowel. If the plosive is voiced (e.g. /b/, /d/, /g/) vocal fold vibration begins concurrent with, or slightly after, the release. For voiceless English plosives (e.g. /p/, /t/, /k/), the start of vocal fold vibration is delayed for some tens of milliseconds (typically more than 40 ms), but aspiration noise (from turbulence in the vocal tract or at the glottis) is usually present in the interval between release and the start of voicing. This interval, known as *voice onset time* (VOT), is the most important articulatory difference between voiced and voiceless plosives. Differences in VOT seem to distinguish voiced from voiceless plosive stops in a wide variety of languages although the absolute values can differ. In Puerto Rican Spanish, for instance, voiceless plosives are characterized by the initiation of vocal fold vibration soon after release of the stop (within 20–25 ms for bilabial /p/) but voiced plosives are significantly prevoiced; vocal fold vibration commences some time before the plosive release (in

citation form, typically over 100 ms in bilabial /b/ (Lisker and Abramson, 1964; Williams, 1977)).

On the other hand, distinguishing the *place of articulation* (e.g. bilabial /b/ from alveolar /d/ from velar /g/, or /p/ from /t/ from /k/) seems to depend primarily on the spectrum of the release burst and the changing resonances of the vocal tract as the stop is released. The acoustic nature of the plosive burst will depend on the shape of the vocal cavities and hence on the place of constriction. Similarly, the dynamic pattern of the changing resonances of the vocal tract as it moves from a constriction to a vowel, shaping the spectrum of voicing or aspiration, will be different for different places of constriction.

Although this short survey makes no claim to be a complete catalogue of the articulatory manoeuvres associated with the speech of the world's languages (or even of any particular accent of English), it does cover some of the most important features of all languages: the presence or absence of vocal fold activity; the frequency of vibration of the vocal folds, if active; and the shape of the vocal tract (including the possibility of nasal coupling and the degree to which constrictions can cause air flow to generate an aperiodic excitation). All of these may vary dynamically. We will now consider the acoustic structure of sounds generated by these mechanisms.

B *Acoustic features*

Most of the primary acoustic features of speech sounds can be satisfactorily explained by the *source-filter theory of speech production* (Fant, 1960). The essential aspect of this theory is that it models speech production, to a first approximation, as being composed of two noninteracting processes—the generation of sound by a *source* whose spectrum is shaped by a linear *filter*. We begin our description of the model with reference to the production of steady-state non-nasalized vowels, a particularly simple case.

Since vowels are normally voiced, the source (often known as the glottal source) is taken to be situated in the larynx, with the sound arising from the quasi-periodic vibrations of the vocal folds. This near periodicity of the glottal waveform gives rise to an essentially harmonic spectrum with energy at the frequency of vocal fold vibration and its integer multiples. Hence, as the vocal fold vibration rate increases, so does the frequency of the fundamental component in the spectrum and the inter-harmonic spacing. As it decreases, the harmonics become more densely packed with the fundamental component going down in frequency. It is found empirically that the spectral envelope of the glottal source typically has a downward slope with increasing frequency of about 12 dB per octave, with little energy above 4 kHz or so, although this can vary with speaking effort, voice quality, and across individuals.

Before the sound from the glottal source can reach a listener, it must pass

through the vocal tract, which can be modelled as a linear filter. Hence, in terms of the present model, the effect of the vocal tract filter on the glottal source can be completely specified by its frequency response or transfer function, which is in turn determined by the shape of the vocal tract. For nonnasalized vowels, when the velum closes off the nasal cavities, the vocal tract has no antiresonances, only resonances. These resonances are called *formants*, and only the lowest four or five are important in speech perception.

To obtain the spectrum of the output of the vocal tract, one simply multiplies the spectrum of the glottal source by the transfer function of the vocal tract filter. The formant resonances of the vocal tract filter give rise (as long as there is energy from the source at that frequency region) to peaks in the output spectrum, also known as formants. Formants are usually identified by numbers from 1 upwards, starting from the formant at the lowest frequency. Finally, the radiation of sound from the mouth can be approximated by a high-frequency emphasis that increases by 6 dB per octave. An idealized example for a particular vocal tract shape and voice fundamental frequency is shown in Fig. 7.4. Note that the harmonic structure of the spectrum arising from the quasi-periodic source is preserved whilst its spectral envelope is modified. These two factors can, as the theory assumes, be altered more or less independently. Changes in the rate of vocal fold vibration (perceived as changes in voice pitch) can be made independently of changes

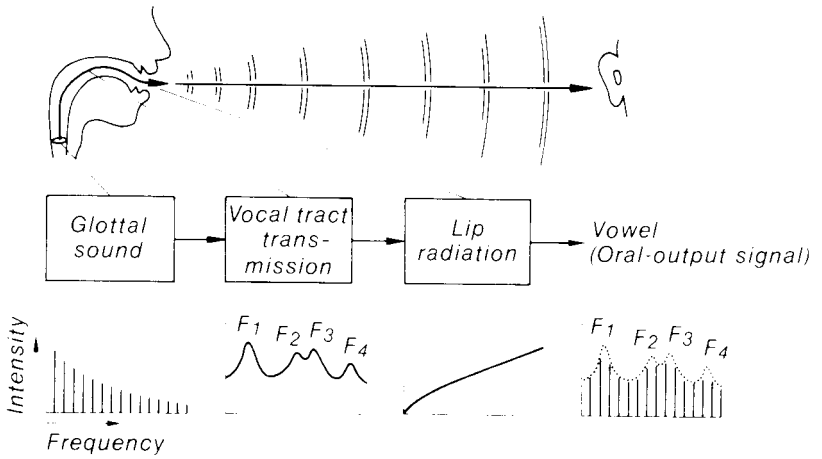


FIG. 7.4 The source-filter theory of speech production applied to the production of a steady-state vowel. Linear scales are used for the frequency axes, and logarithmic ones for the intensity axes. The periodic waveform arising from laryngeal activity gives rise to a harmonic spectrum with an overall spectral envelope that decreases with increasing frequency (bottom, far left). The vocal tract resonances, as well as the high frequency emphasis due to radiation effects, impose further modifications on the spectral envelope, but the harmonic nature of the source spectrum is preserved. Reproduced with permission from Hirano (1981).

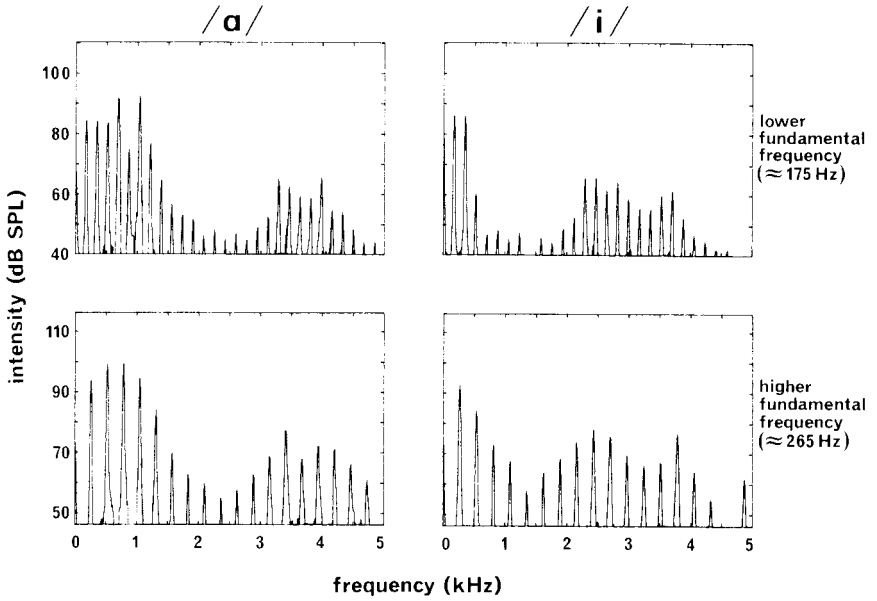


FIG. 7.5 The spectra of two steady-state vowels being uttered at two different fundamental frequencies. The fundamental frequency determines the inter-harmonic spacing, while the vowel quality determines the spectral envelope. /ɑ/ is characterized by a high F_1 and a low F_2, while /i/ is characterized by a low F_1 and a high F_2 (see Fig. 7.22 for typical formant relationships among a complete set of English vowels). Note how the two closely spaced lower formants of /ɑ/ cannot be resolved with the relatively large inter-harmonic spacing of the higher fundamental frequency. All spectra were obtained from continuously held vowels uttered by a single male speaker, using a time window of 80 ms.

in the shape of the vocal tract (perceived as changes in vowel quality). To a useful first approximation, the source and filter can be treated independently both acoustically and perceptually. Figure 7.5 shows the combination of two different vocal tract shapes excited by the vocal folds vibrating at two different rates.*

The acoustic structure of nasalized vowels and voiced nasals can be modelled in the same way, except that now antiresonances, or zeroes, can occur in the filter function. These give rise to valleys in the output spectrum sometimes called *antiformants*. The essential property of the model, in which a harmonic spectrum is shaped in spectral content, is preserved (see Fig. 7.6 for examples).

Voiceless sounds require the use of a different *source*—one that is aperiodic, and may be considered, in auditory terms, to consist of a broad

*It is not strictly true that the source and filter operate independently. The opening and closing of the vocal folds introduces time-varying subglottal resonances. Similarly, the load presented to the glottal source by the vocal tract varies as its shape changes. For further discussions, see Koizumi *et al.* (1985) and Ananthapadmanabha and Fant (1982).

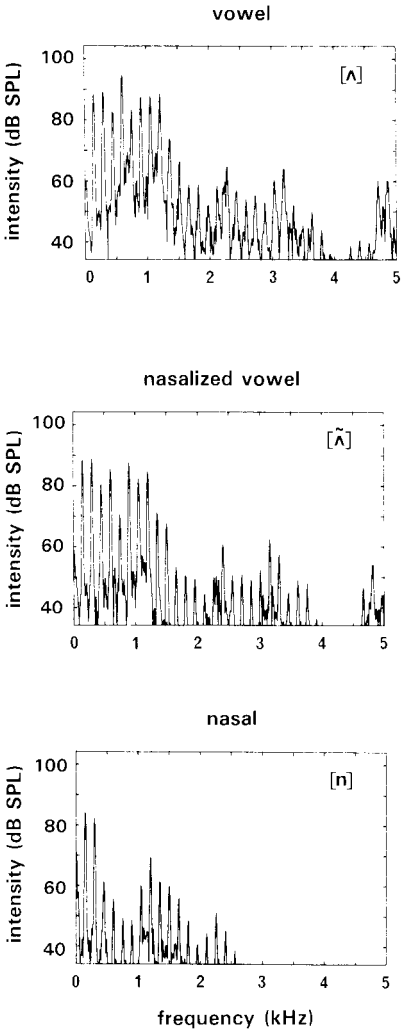


FIG. 7.6 The spectra of three sounds which can be modelled as resulting from a quasi-periodic source being modified by a vocal-tract filter with resonances, and sometimes antiresonances. The [ʌ] as in 'pl*u*ck' (top) is produced with the nasal cavities isolated from the vocal tract by the closed velum; hence only resonances are present. In the middle spectrum, the velum of the speaker has been dropped to create a nasalized version of the [ʌ], symbolized as [ʌ̃]. Here, the amplitude of the fourth and fifth harmonics has been drastically reduced, presumably as a result of an antiresonance of the nasal cavities between these two harmonics. Finally, on the bottom, is the spectrum of the nasal murmur of [n] (i.e. the part of the sound that can be sustained). Not only is there probably an antiresonance in the region of the fifth to sixth harmonics, but also all the harmonic components above about 2 kHz have been attenuated. Hence the spectrum is much more heavily weighted to the low frequencies. All three spectra were obtained from continuously held sounds uttered by a single male speaker with a fundamental frequency of about 150 Hz, using a time window of 80 ms.

band of frequencies. Again the source (now wide band and located at the point of constriction) is filtered (by the cavities in front of and behind the constriction) and radiated to obtain the output spectrum, which will now be continuous. Different places of articulation (points of constriction) effectively alter the properties of the source as well as those of the filter, but the final result can always be described as a wideband noise with the appropriate spectral shaping. Figure 7.7 shows some typical spectra for two English voiceless fricatives, /s/ and /ʃ/.

A voiceless source may also be transient, as happens in the release burst of plosive stop consonants, rather than continuous as for fricatives. Again, different places of articulation cause differences in the frequency content of the acoustic signal.

Modelling the production of voiced fricatives (like /z/ in *z*ero and /ʒ/ in plea*s*ure) requires the use of a *mixed* source in which both periodic energy (due to the vibration of the vocal folds) and aperiodic energy (due to turbulence noise at a constriction) occur together. This will lead to a spectrum that has both continuous and harmonic components.

Finally, dynamic aspects are accounted for by allowing the source, and/or the filter function to vary in time. Changes in voice fundamental frequency necessitate changes only in the rate of the periodic source representing voicing. Diphthongs at a constant voice pitch are modelled as relatively slow changes in the filter function alone.

Graphically portraying the spectra of such dynamically varying sounds requires the addition of the time dimension to the two of amplitude and frequency we have been using. One particularly useful way of doing

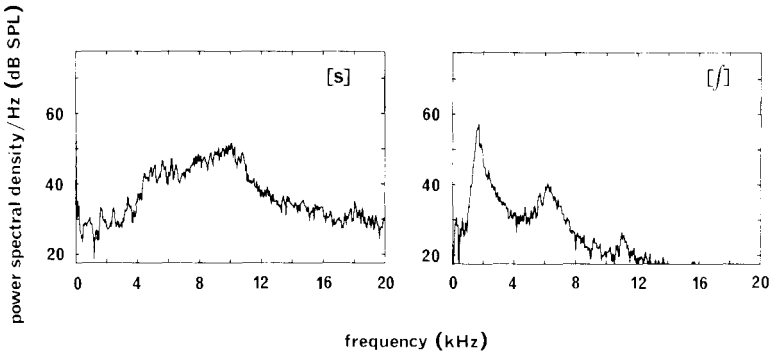


FIG. 7.7 The spectra of two continuously held voiceless fricatives uttered by a single male speaker. Note that as the sounds are not periodic, continuous spectra are obtained. The [s] as in '*s*eep' (on the left) has its energy concentrated in a rather higher frequency region than the [ʃ] of '*sh*eep'. Both spectra can be considered to result from a white-noise source being appropriately filtered. Each spectrum is the result of averaging 12–14 different spectra spaced through a single continuously held utterance, each spectrum obtained with a time window of 20 ms to obtain a spectral resolution of approximately 50 Hz.

this is incorporated in a sound spectrogram, a graphical display of the result of passing a sound through a bank of bandpass filters with closely spaced centre frequencies. The degree of blackness of a trace is used to indicate amplitude, reserving the abscissa and ordinate for time and frequency respectively. Figure 7.8 shows a spectrogram of the diphthong /ɒɪ/—as in 'Oi!', a working-class London English equivalent of 'Hey!'—and steady-state versions of the two vowel qualities it encompasses. Note the presence of the formants which are clearly visible as dark bands.

Since these sounds are voiced, the quasi-periodic excitation of the larynx source is reflected in the quasi-periodic vertical striations. Each striation represents the vocal fold closure in one cycle of vibration. Hence the instantaneous voice fundamental frequency can be determined by taking the reciprocal of the time between two successive striations. Individual harmonics of the voice source are not visible here because a relatively wideband (300 Hz) analysis was used. This smears spectral detail. Formants, being typically rather farther apart in frequency than the harmonics of the voice source, are usually resolved, and the wide bandwidths of both the formants and analysis filters preserve the temporal detail necessary to resolve each vocal fold oscillation.

Again, the plosive stops can serve as dynamic sounds *par excellence*, in which both the filter and the source functions must change appropriately in time. Figure 7.9 shows spectrograms of the three English voiced/voiceless plosive

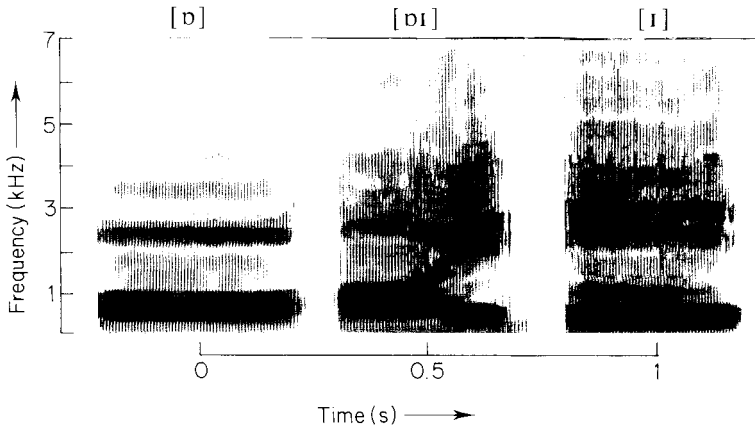
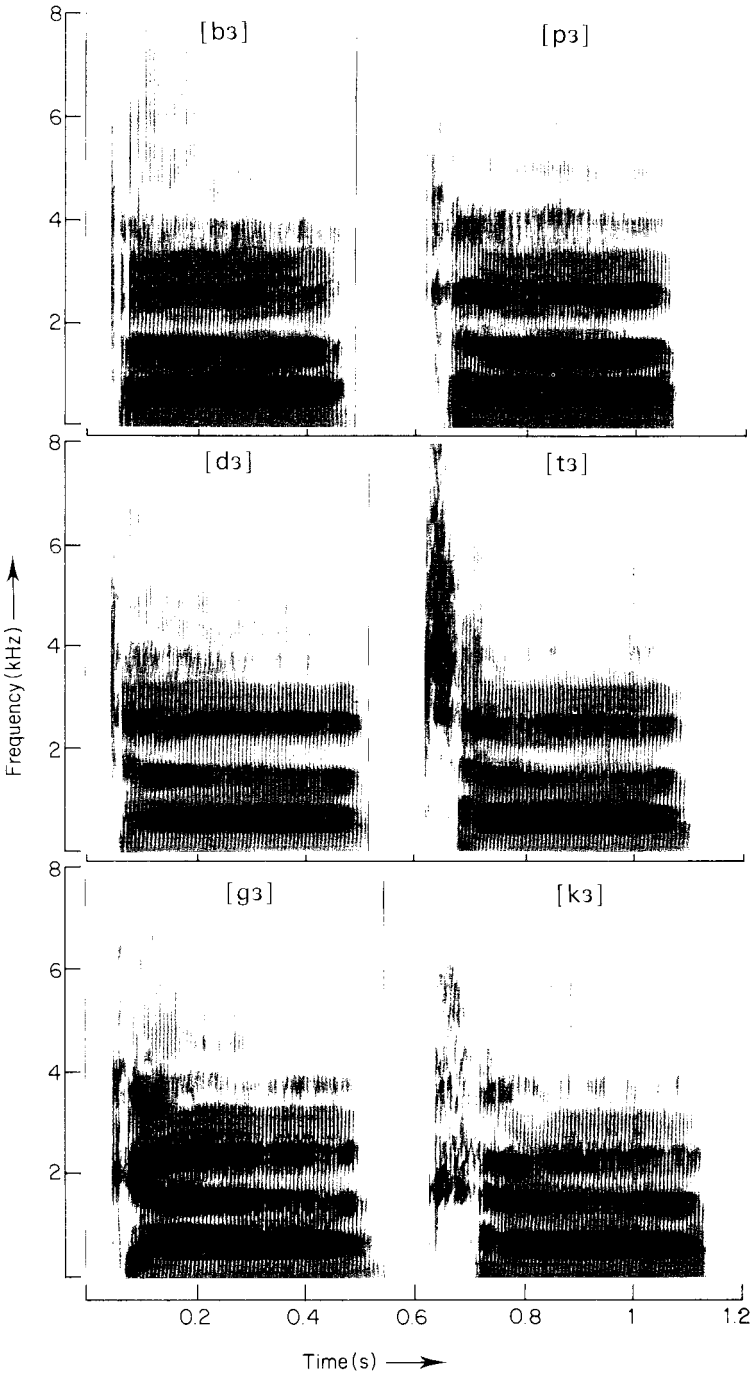


FIG. 7.8 Spectrograms of a diphthong and two steady-state vowels which have a quality similar to the initial and final qualities of the diphthong. All the sounds were from a speaker with vowels roughly appropriate for British 'Received Pronunciation' (RP—otherwise known as 'the Queen's English'). [ɒ], at left, has its first two formants too close together to be resolved in the spectogram. Similarly, in [ɒɪ] (centre), the fact that the lowest dark band contains two formants does not become clear until about halfway through the sound, when they begin to separate. F_1 moves to a slightly lower frequency, while F_2 moves to a considerably higher one, these positions being appropriate for the vowel [ɪ], as seen at the right of the figure. (Fig. 7.22 shows for RP vowels the typical relationships between the first two formants.)

pairs in initial position. In all the syllables, we see a transient burst, due to the sudden rush of air from the release of the pressure built up during vocal tract closure. The spectrum of this burst is determined primarily by vocal tract resonances and antiresonances at release. In this release phase of the plosive production, the energy source is *transient*. For voiced plosives (/bɜ/ 'buh', /dɜ/ 'duh' or /gɜ/ 'guh'), voicing starts almost immediately thereafter, as seen in the striations reflecting quasi-periodic vocal fold vibration. The source has changed both in position along the vocal tract, and from being aperiodic to being voiced. Through the burst and onset of voicing, the vocal tract is changing shape, as it moves from closure to the position appropriate for the following vowel. Thus the filter is changing, and this is reflected in the formant transitions in the spectrograms. Different places of articulation give rise to different dynamic filter functions and hence to different dynamic formant patterns.

For the same place of articulation, the variation of the vocal tract filter with time is similar for voiceless and voiced plosives. In our source-filter description, the crucial distinction between voiced and voiceless plosives is in the way the source changes. Following the transient excitation of the release burst, air continues to flow, but in the voiceless consonant the vocal folds are initially tensed so that they do not begin to vibrate for some tens of milliseconds (delimiting VOT). In this interval, the source remains aperiodic (although typically at a lower intensity than at the transient release), due to turbulence in the vocal tract and especially in the glottis. The resulting sound is known as *aspiration* because it has an [h]-like character. Since this aspiration noise is shaped by the vocal tract filter, formant transitions are still present in the spectrograms. Now, however, since there is no quasi-periodic excitation, no striations are seen. The aperiodic excitation shows itself as a random pattern of spots of different degrees of darkness. The upper formants are revealed rather more clearly during aspiration than F_1, which is usually invisible, because the spectrum of the aspiration source is, in comparison to the voicing spectrum, rather less intense in the low frequencies than in the high. This difference arises from the manner of its generation and from the absorption of low frequency energy into the subglottal system through the relatively open glottis. Furthermore, the spectrogram itself adds its own high-frequency emphasis, at least partially to mimic the normal variation of auditory thresholds with frequency.

In summary, the timing of source changes is important for distinguishing voiced from voiceless plosives, while the dynamic changes of the filter function distinguish different places of articulation. This relatively simple description gives rise, through the source-filter model, to a number of additional associated acoustic features. The resonance associated with F_1, which always rises from a low position as the initially closed vocal tract moves towards the position appropriate for the following vowel, is not normally


[b3]
[p3]
[d3]
[t3]
[g3]
[k3]
8
6
4
2
Frequency (kHz)
0.2
0.4
0.6
0.8
1
1.2
Time (s)

substantially excited by the aperiodic aspiration source, and hence is not perceptually important until voicing starts. Often, the F_1 transition is complete before the onset of voicing in voiceless plosives, and so they tend to be characterized by an F_1 that is relatively flat. Voiced plosives usually have an F_1 that is rising relatively rapidly at the onset of voicing. For the same reason, the onset frequency of F_1 covaries with VOT (the longer the VOT, the higher the onset). Such acoustic features are clearly not due to either the source or filter on their own, but to an interaction between the two.

Just as for our summary of articulatory processes, we do not pretend to have described exhaustively the acoustic aspects of speech. On the other hand, the features described are certainly among the most important in the majority of the world's languages: whether the spectrum is continuous or harmonic (or both); the spacing of the harmonics if present; and spectral shaping (including overall amplitude). All these may vary dynamically. Figure 7.10, a spectrogram of the phrase 'frequency selectivity', shows how one short utterance can illustrate most of these possibilities.

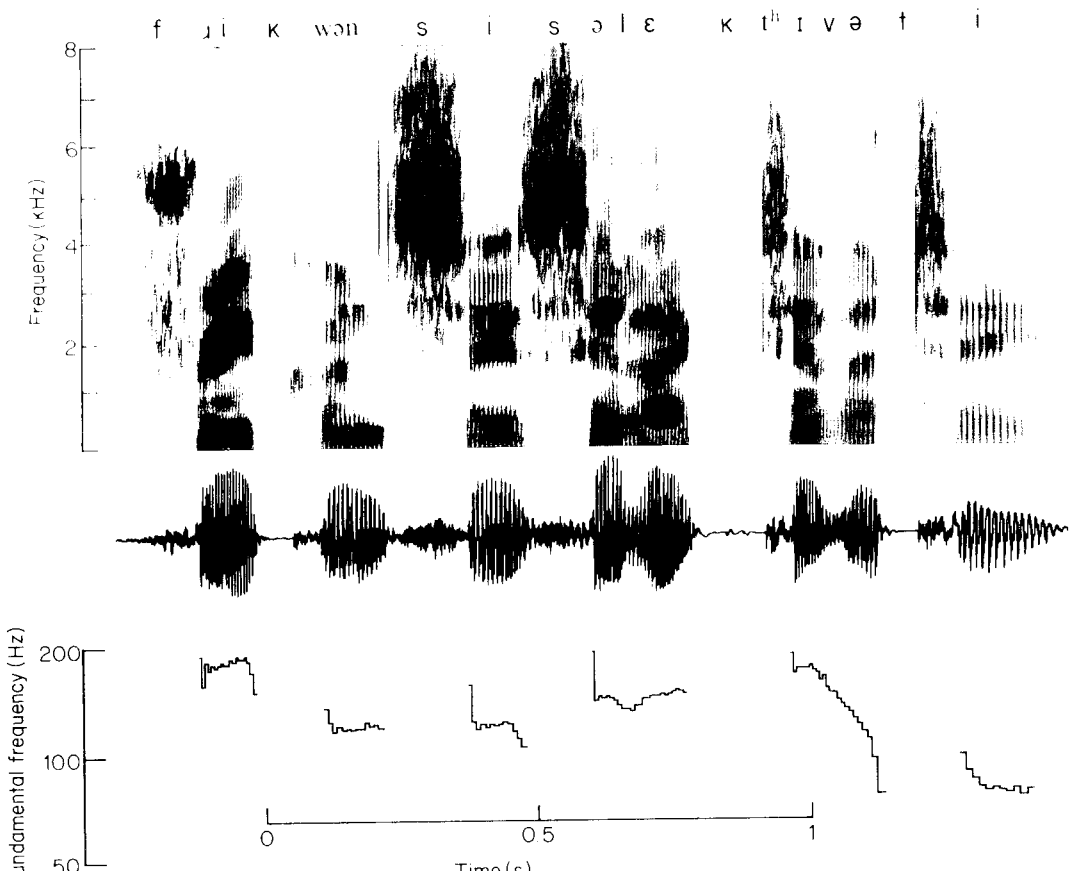
C *Three caveats*

These summaries of the functional anatomy and acoustic characteristics of speech will serve us adequately in our discussions of how auditory frequency analysis might affect speech perception. There are, however, three important issues that we have either skirted or avoided.

Firstly, although we have talked about some speech sounds that are inherently dynamic, we have discussed others (like vowels and fricatives) as if they were completely static in spectral characteristic, a phenomenon rare outside the laboratory. In free-running natural speech, things are always changing. Even vowels, supposedly the prototypic steady-state sounds, rarely show unchanging spectra over anything but the shortest stretches of time. Also, when uttered in natural context, the formants (understood to be a main cue of vowel identity) often do not reach the values they attain when uttered in citation form, a phenomenon known as undershoot (for a review see Miller, 1981).

Secondly, the acoustic characteristics of a particular speech sound are determined in part by the sounds around it. For example, the word 'team' has an aspirated /t/ (i.e. a puff of air following the plosive release), and 'ream' has a voiced /r/; but in 'steam' the /t/ has no aspiration, and in 'tree', the /r/ is unvoiced. This is known as coarticulation, and it operates to a greater or lesser extent for all speech sounds. There is thus no isomorphism

FIG. 7.9 *(opposite)* Spectrograms of the six English plosives spoken before the same vowel [ɜ] 'uh' from a British English speaker. For further details, see the text.


f ɹ i k wɑn s i s ə l ɛ k tʰ ɪ v ə t i
8
6
4
2
Frequency (kHz)
200
100
50
Fundamental frequency (Hz)
0
0.5
1
Time (s)

between acoustic features and the phonemes they specify. A specific acoustic feature can signal different phonemes in different circumstances, and the same phoneme can be signalled by acoustically distinct patterns, depending on the context. Even sounds that are relatively stable in their acoustic

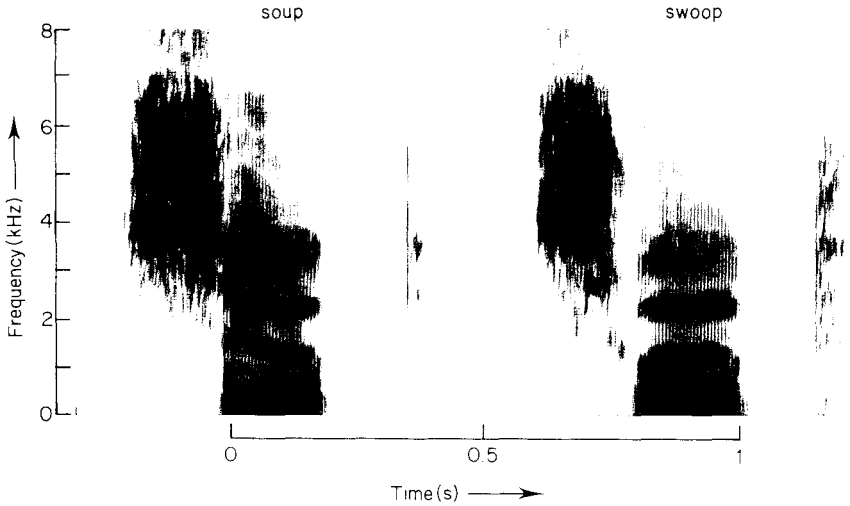


FIG. 7.11 Spectrograms of the words 'soup' and 'swoop' as spoken by a male speaker of British English. The [s] in 'soup' is relatively steady-state in spectral shape through its total duration. In 'swoop', however, the low frequency edge of the band of noise associated with the [s] 'swoops' downward, as the resonances of the vocal tract are lowered by the preparatory liprounding (effectively lengthening the vocal tract) being made for the following [w]. This is an example of anticipatory coarticulation. Based on an example and figure in Pickett (1982).

FIG. 7.10 (*opposite*) A spectrogram of the phrase 'frequency selectivity' [fɹikwənsi səlɛktʰɪvəti], as spoken by an adult English male, together with its pressure/time waveform (middle trace) and fundamental frequency contour (at bottom). Voiced segments are indicated by: (1) vertical striations in the spectrograms; (2) intervals of relatively high-amplitude quasi-periodic traces in the time waveform; and (3) the presence of a trace in the fundamental frequency contour. Relatively high larynx frequencies (as in the vowel [i] of 'frequency') are indicated by closely spaced striations on the spectrogram, while low larynx frequencies (as on the final syllable) have their striations further apart in time. Note the presence of turbulence noise in the high frequencies associated with the voiceless fricatives [f] and [s], and the aspiration of the voiceless plosives [k] and [t] (although less intense for [k]). Since there is no voicing here, no fundamental frequency contour exists over those intervals. The pressure/time waveform does show activity, however, in the form of a relatively low level aperiodic waveform. Interestingly, [v] has little or no frication associated with it, but is clearly voiced. The nasal [n] shows the expected antiresonance in the spectrogram as a large white area in the region of 1.5–2.0 kHz. Dynamic changes in voiced segments are clearly seen in [ɹi] and [əlɛ]. The spectrogram was made using a differentiating equalizing filter, equivalent to a frequency response with a characteristic that rises at 6 dB/octave. The time waveforms were plotted after low-pass filtering at 7 kHz and digitization at 14.3 kHz. The fundamental frequency contours were derived by a period-by-period analysis of a laryngographic recording (Fourcin, 1981) – this accounts for the 'step-like' appearance of the traces.

characteristics across phonetic environments (like /s/) can show coarticulatory effects. Figure 7.11 shows a particular example.

Finally, a point related to the previous one: all speech sound contrasts can be based on a subset from a *number* of acoustic cues. The specific cues used and their relative importance vary with the particular context, and across listeners. For example, the distinction between /tʃ/ (as in dit*ch*) and /ʃ/ (as in di*sh*) is known to be influenced by, at least: the rise time and duration of the frication, the presence or absence of a release burst, the presence or absence of a silent gap after the preceeding vowel, and the properties of that vowel (Gerstman, 1957; Dorman *et al.*, 1980). Accounting for the ways in which human listeners appropriately integrate these cues into a phonemic decision is a central difficulty for any theory of speech perception, and is currently the subject of much speculation (e.g. Liberman, 1982; Pastore, 1981; Repp, 1983). It is enough for our purposes simply to realize how many different factors can enter into a decision on the identity of a particular speech sound.

To summarize the three caveats:

1 Speech sounds are, on the whole, dynamic in nature.
2 There is no simple one-to-one correspondence between acoustic structure and phonemic class.
3 All speech sound contrasts depend on a number of interacting acoustic cues.

The last two of these will have little influence on our discussions, serving only to caution against the belief that there is a simple relationship between the physical structure of speech sounds and their phonemic identity. As long as we deal with the perception of specific *acoustic* features (e.g. formant transitions), and their possible roles, as opposed to making generalized statements about the perception of phonemic contrasts (e.g. telling /d/ from /g/), there will be little trouble.

The first caveat, on the other hand, is more problematic. For example, the great bulk of the work on the perception of vowel quality (as well as the related psychophysical investigations of timbre) has been done using steady-state stimuli. Although hardly realistic, these are the only data available. However, it seems likely that much of what is learned from such stimuli is applicable at least to the relatively slowly changing natural vowel spectra. It is probably not directly applicable to quickly changing spectra like those associated with plosive consonants.

3 THE ROLE OF FREQUENCY SELECTIVITY IN NORMAL AND IMPAIRED SPEECH PERCEPTION

Having reviewed the major articulatory and acoustic characteristics of various classes of speech sounds, we now turn to our main topic, the role of

frequency selectivity in their perception. Again, our discussions must be 'selective', and we shall examine in detail only those acoustic contrasts relevant to speech that seem heavily dependent on auditory frequency analysis. So, for example, although the presence or absence of vocal fold vibration may well be the most basic contrast of all in speech, this distinction probably relies relatively little on the properties of auditory filters. Hence, we begin our discussions with another crucial aspect of the activity of the vocal folds—their rate of vibration.

A *Static and dynamic fundamental frequency patterns*

As our writing system is based on the indication of specific individual speech sound contrasts, we tend to forget the importance of features that are either not explicitly represented, or at best indicated only very crudely in our script. The accent differences associated with different vowel systems are not catered for by the text on this page, for example (luckily for the two authors who use different vowel systems), and voice pitch, length and loudness are even less well notated. Yet attempts to synthesize speech without due attention to these latter *prosodic* features lead to unnatural and unacceptable speech. Here we shall only discuss the role of voice fundamental frequency, both because of its linguistic and developmental importance, and because its perception reflects the operation of the mechanisms of frequency selectivity.

i Importance in speech contrasts

As we discussed above, voice fundamental frequency is determined by the time between successive closures of the vibrating vocal folds. Since this vibration is generally nearly periodic, it leads to a sensation of pitch which finds its purest expression in singing. Singers control the rate at which their vocal folds vibrate to convey a particular temporal pattern of pitch variations—a melody. But all speakers of all languages produce sophisticated speech melodic contrasts to convey messages, although the form these take may differ from language to language.

The information conveyed by voice pitch changes is of three basic types, two of which—paralinguistic and intonational—are common to all languages.

Information about the sex, age, identity and emotional state of the speaker is known as paralinguistic since it is not determined by the linguistic (phonemic, lexical and syntactic) structures which determine meaning. So, for instance, the average voice fundamental frequency of women is higher than that of men, and is higher in children than in adults. Also, we are all well aware of how others can signal boredom, interest, anger or excitement (without words to that effect) by their manipulation of voice pitch contours.

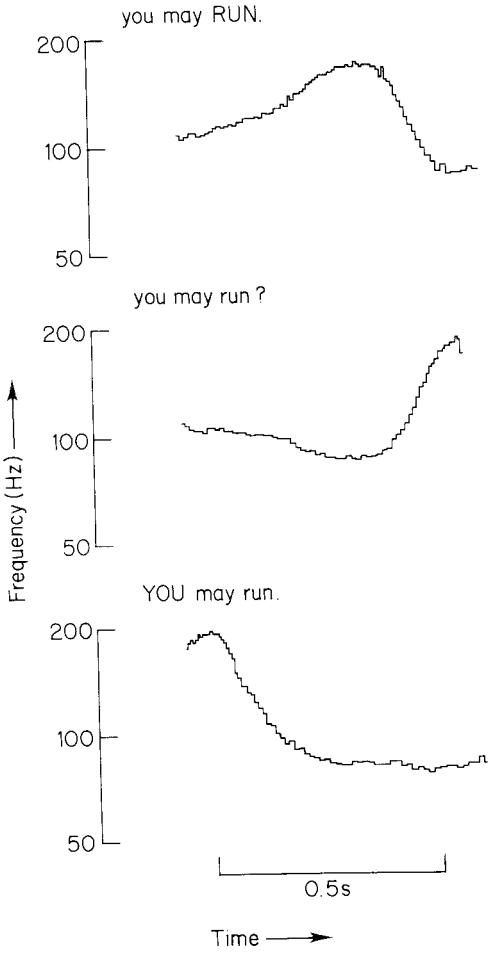


FIG. 7.12 Contrastive fundamental frequency contours derived from a period-by-period analysis of laryngographic recordings from an adult male speaker (Fourcin, 1981). In each case, the continuously voiced utterance was 'you may run'. In the contour at the top of the figure, the word 'RUN' was stressed, while 'YOU' was stressed for the contour at bottom. Both of these were uttered as declarative statements, as opposed to the question form of the middle example. As all these utterances were spoken in citation form, the contours extend over a rather larger frequency range than would be found in conversational speech.

Intonation refers essentially to those changes in voice fundamental frequency that are related to the syntactic structures of an utterance, but which do not change the dictionary meaning of individual words. In this sense, intonation functions as punctuation does in written language, but is much richer. For example, voice pitch changes can be used to mark syntactic units, to clarify ambiguous pronoun references, to indicate important words,

TABLE I
An example of the use of tones in Cantonese. The utterance /ji/ means six different things depending upon the voice pitch contour used by the speaker (from Ching, 1981)

Voice pitch contour	Meaning
high level	clothes
high rising	chair
mid level	meaning
low falling	child
low rising	ear
low level	two

or to distinguish questions from statements. Figure 7.12 shows examples of the latter two.

Many other languages have yet a third function of voice pitch changes found only rarely in English, or the other European languages.* In these *tone* languages, changes in fundamental frequency are used to indicate changes in the dictionary meaning of a word. For example, in Cantonese Chinese, the unglottalized monosyllable /ji/ (pronounced 'yee') can have six different unrelated meanings, depending on the voice pitch contour used (see Table I). Tones can also have a grammatical function, indicating verb tense or noun case (Ladefoged, 1975).

In addition to these important types of information that voice pitch provides to the normally hearing adult listener, it also seems to serve a special role both for the developing infant, and for the hearing impaired. Although there is much controversy regarding the acoustic patterns to which infants are most sensitive (Fourcin, 1978; Kuhl, 1976), there is little doubt that, productively, contrastive control of voice fundamental frequency precedes that of any other feature (Ricks, 1975).

For the hearing-impaired listener who comes to rely on lipreading, voice pitch information is crucial, as it is relatively invisible. Some of the important cues to the identity of individual sounds depend only on detecting the presence or absence of vocal fold activity; of more importance seems the appreciation of variations in vocal fold vibration rate. Rosen *et al.* (1980) have shown in a connected discourse tracking task that although the presence or absence of voicing can be a significant help to lipreading, the variations in voice fundamental frequency are of greater importance, even to speakers of English.

*There are a few cases in European languages in which fundamental frequency contours can cue differences in the dictionary meaning of a word. In English, for example, differences in fundamental frequency contour can cue the difference between the noun *sub*ject and the verb sub*ject* (italics indicating the stressed syllable; Fry, 1958), and between other similar pairs (e.g. permit, contract). For a review, see Fry (1968).

For Cantonese, Ching (1981) has shown a similar result. Information about the presence or absence of voicing improved lipreading to about the same degree as it did in English, but variations in voice pitch were of even greater help than they were in English. This is hardly surprising given the lexical role of voice fundamental frequency in Cantonese.

ii Major acoustic features

We are always, when considering speech, dealing with the pitch of complex tones, never single (or even a small number of) sinusoids. Voiced speech always has a strong fundamental (except in artificial situations like using a telephone) and a rich harmonic spectrum. Therefore, unlike the case with many of the unnatural stimuli used in psychoacoustic investigations, the perceived pitch of any voiced speech sound is usually highly salient and nonambiguous. Exceptions to this rule occur, of course, in pathological conditions, and even in normal speech. In particular, during creaky voice (typically known as vocal 'fry' in the USA), common in many dialects of English (especially at the end of voice pitch falls), vocal fold vibrations are low frequency and irregular. This can give a 'rough' or diplophonic (two-pitched) quality to the speech.

The range of fundamental frequencies found in natural speech is rather

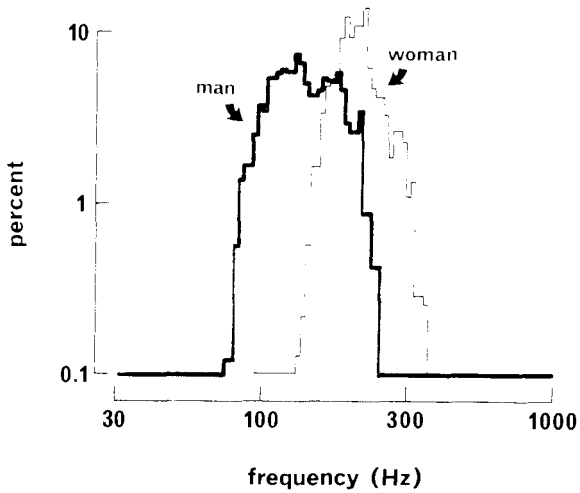


FIG. 7.13 Histograms of the distribution of voice fundamental frequency for two adult speakers, derived from laryngograph recordings of the reading of a prepared text. The duration of the recordings used in the analysis was two minutes for the man and five minutes for the woman. These histograms were constructed on a period-by-period basis in which fundamental frequency was quantized to 100 logarithmically-spaced values between 10 Hz and 1 kHz. For further details of the analysis procedure, see Fourcin (1981). The male speaker used here was the same as that used for the radiographs of Fig. 7.2, and whose utterances are analysed in Figs 7.8, 7.9, 7.10, 7.11 and 7.12.

restricted compared to that found in music. While the fundamental frequency of a piano extends from 27 Hz to 4186 Hz, voice fundamentals are contained in the region between about 70 and 1000 Hz. No particular speaker will exhibit this range, of course, the maximum being about two octaves for any particular adult speaker. Even this range is not likely to be expressed in ordinary speech. Figure 7.13 shows some examples of histograms of voice fundamental frequency which occurred when an adult male and female read a standard passage.

It can also be seen in Fig. 7.13 that average voice fundamentals differ across speakers. Adult males as a class show the lowest mean frequency (about 100–130 Hz) and young children the highest (about 300 Hz at 10 years of age). Adult females fall between with means of about 190–220 Hz (Fant, 1956; Hirano, 1981).

iii The performance of normal listeners

Although much of the psychophysical data regarding the discriminability of frequency changes is not directly applicable to the case of speech, some generalizations can be made. Perhaps most importantly, frequency discrimination of complex tones is better than that for sinusoids at the fundamental, for fundamental frequencies in the voice pitch range. In other words, adding harmonic components leads to smaller just-noticeable differences (jnds) in fundamental frequency (Flanagan and Saslow, 1958; Henning and Grosberg, 1968; Fastl and Weinberger, 1981; Stock and Rosen, 1986).

Henning and Grosberg claimed that this improvement was due to listeners using the particular harmonic in the complex (one near 2 kHz) that gave best performance. As all harmonics in a periodic complex change by the same proportional amount for a given change in fundamental, and the jnd for sinusoids in percentage terms has a broad minimum near 1–2 kHz (Harris, 1952; Wier *et al.*, 1977), listeners will do best by directing their attention to that frequency region. This approach assumes that the ability to hear frequency changes in a sinusoid is the same whether the tone is presented on its own or as part of a complex.

In fact, this is not the case, as shown in a study of the discriminability of individual harmonics within an otherwise fixed complex by Moore *et al.* (1984). Although these experimenters did not make a direct comparison of sinusoidal frequency discrimination performance in isolation and in the complex, all the listeners were highly trained, so it is possible to compare their results with those obtained by Wier *et al.* (1977) for isolated sinusoids. We shall only look at the results obtained with a complex tone consisting of 12 equal amplitude harmonics of a 200-Hz fundamental.

For about the first three harmonics (200–600 Hz), performance with the harmonics in the complex was roughly equivalent to that obtained for sinusoids in isolation (0.25–0.3%). For the rest of the harmonics, however (except the highest, a special case we ignore as it is not relevant to speech),

performance for the harmonics within the complex was considerably worse than the approximately 0.2% obtained for sinusoids, reaching 2 to 5% for harmonics 9 to 11. Finally, discrimination was more acute for the entire complex (0.13–0.22%) than for any single harmonic. This means that Henning and Grosberg's hypothesis is wrong in two ways. Not only is information combined across harmonics, but it is the low-frequency harmonics, *not* those in the 1–2 kHz region, which primarily determine the discriminability of the complex as a whole, at least for complexes with equal-amplitude harmonics. Interestingly with regard to speech, with its well-defined formant structure, Moore *et al.* found that making one harmonic more intense than those around it increases its discriminability and hence makes it more important in determining the discriminability of the entire complex.

This ability to combine information across harmonics results in very fine discrimination indeed. For example, Stock and Rosen's (1986) better listener gave jnds of about 0.3% over the range 125–500 Hz for trains of 100 μs pulses. Henning and Grosberg found a jnd of about 0.2% at 250 Hz. Furthermore, as would be expected, these values are similar to those found with steady-state vowel sounds. Flanagan and Saslow (1958) and Klatt (1973) report jnds of about 0.2 to 0.5% for vowel sounds with fundamentals of 80 and 120 Hz.

Does this mean that listeners are normally sensitive to such small changes in running speech? Of course not. First of all, it is highly unlikely that barely perceptible differences (only 75% detectable) would be used to convey information. Secondly, the stimuli used in these experiments are hardly representative of natural speech, in which neither the spectral envelope nor the fundamental frequency is static for very long. Therefore, it is important to know how variations in either affect frequency jnds if implications for speech perception are to be drawn.

Klatt's (1973) experiments addressed this issue explicitly. He measured jnds for an unchanging vowel with flat fundamental frequency contours, and also with a series of simple linear rising and falling contours around 120 Hz. When both contours were falling at the same rate (30 Hz in the 250 ms duration of each stimulus) the jnd increased to 2 Hz, some 7 times the jnd of 0.3 Hz found for a steady-state contour. When the contours differed in their slopes (again with the standard decreasing 30 Hz in 250 ms, but with the comparison stimulus changing by a greater amount), the just-discriminable stimulus had a rate of decrease 32 Hz/s steeper than the standard of 120 Hz/s. As the two contours had the same fundamental exactly half-way through their duration, this translates into a +4 Hz difference at stimulus onset and a −4 Hz difference at offset.

In contrast to these results, using a glide-vowel combination (/jɑ/ as in the American pronunciation of *ya*cht) in place of the steady-state vowel had relatively little effect on the jnds, increasing them by about 25–65%. This is perhaps not too surprising as the /jɑ/ was steady-state for roughly half its duration. Jnds may be yet larger

for sounds with spectral variation over the entire course of their existence.

Pierrehumbert (1979), in the context of an investigation into the perceptual aspects of fundamental frequency declination (the tendency of voice pitch to be lowered over the course of an intonation group), presents data that may be used to estimate jnds for fundamental frequency in a rather more speech-like context. Her stimuli were created from a natural nonsense utterance 'ma-MA-ma-ma-MA-ma' (stressed syllables in upper case) in which the speaker attempted to preserve the prosodic pattern of 'The baker made bagels'—clearly an East Coast American obsession. This stress pattern led to two peaks in the fundamental frequency contour, one on each stressed syllable. Pierrehumbert modified the fundamental frequency value of the second peak in the utterance, making a series of utterances in which the second peak varied from lower to higher than the first peak. These were then presented to subjects who were required to indicate which of the two peaks was higher in pitch. For first-peak values of 121 and 151 Hz, about a 12 Hz increase was necessary in order to go from 50% 'the-second-peak-is-higher' judgments (i.e. 'they are the same pitch') to 75% 'the-second-peak-is-higher' judgments (which is another way of saying that the jnd is 12 Hz—although the rate of change of the fundamental is also altered here).

Finally, 't Hart (1981) has explicitly attempted to estimate, with yet more natural stimuli, the accuracy with which pitch movements in synthetic speech must be represented to preserve naturalness. Using computer-processed natural four-syllable Dutch number names, he asked subjects to determine which of two pitch movements (both rising or both falling) had the greater extent, and concluded that in no case were the jnds smaller than 6% (or one semitone—about 7 Hz for fundamentals of 120 Hz). Noting that jnd-size changes are unsuitable for communicative purposes, he went on to argue that a three-semitone difference (19% or about 23 Hz at 120 Hz) was usually necessary for two successive pitch movements in the same direction to be reliably heard as different. This is about two orders of magnitude bigger than the 0.3 Hz estimated with steady-state stimuli.

Clearly, experiments which are executed in the psychophysical tradition grossly overestimate the degree to which changes in fundamental frequency are used linguistically. It seems likely that this disparity is yet another reflection of the way in which speech is robust, using only acoustic contrasts which are highly discriminable.* In what way, though, are

*It is interesting to note that this statement is precisely opposite to what was said by the early motor theorists, who proposed that the processes of speech perception were somehow referred back to the processes of production. They argued that a given acoustic contrast was much more likely to be perceived when it was part of a phonemic contrast (Liberman *et al.*, 1961a,b). Part of the better discrimination performance they obtained for speech sounds may be a result of their greater familiarity; the listeners were untrained and relatively inexperienced. There are also questions regarding the appropriateness of some of the nonspeech stimuli they considered as controls for the speech sounds.

any of these properties dependent upon the frequency resolving power of the ear?

iv Modelling voice pitch perception

The theory of pitch perception for complex tones that Moore and Glasberg have detailed in Chapter 5 is directly applicable to the perception of the pitch of voiced speech sounds. Basically, perceiving the pitch of complex tones is viewed as a primarily temporally-based process preceded by an initial frequency analysis. Thus, each temporal processing channel operates on the information contained in a selected frequency region.

The working of this model is best illustrated by examining the output of selected hypothetical auditory filters (each centred on a single harmonic) for a typical voiced speech sound, the vowel [i] with a fundamental of 125 Hz (Figs. 7.14 and 7.15). Here, auditory frequency analysis is modelled as a bank of linear time-invariant filters with appropriate bandwidths.

Generally speaking, at low frequencies, where the filters have the narrowest bandwidths, each filter responds primarily to a single harmonic since the harmonic spacing (equal to the fundamental frequency) is wide compared to the filter bandwidths. There, the temporal information in the waveform is extremely simple and well defined by the peaks (or valleys), which occur at

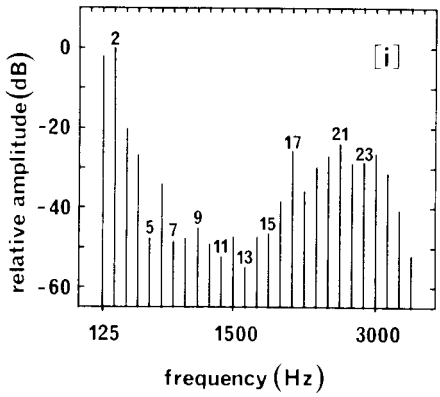


FIG. 7.14 The amplitude spectrum of a synthetic [i]-type vowel modelled closely on a continuously held natural utterance from an adult male. The numbers within the figure indicate the harmonic number of the component with which they are associated. Note how the fifth harmonic is of relatively low intensity in comparison to its immediate neighbours. As antiresonances in this frequency region would not be expected for such a non-nasalized vowel, this variation is almost certainly due to the shape of the source spectrum, which often departs from the idealized 12 dB/octave decrease with frequency typically assumed.

intervals which are integer multiples of the fundamental period. A mechanism that collates information across channels by looking for common intervals would always calculate the fundamental correctly.

On the other hand, high-frequency channels tend to give a rather more complex waveform, since they are being excited by a number of harmonics at the same time. Amplitude modulations of a high-frequency carrier at a rate corresponding to the fundamental can often be seen. A simple peak-picking device would give a whole range of interpeak intervals, some related to the carrier, some related to the fundamental, and some to the time between multiple envelope peaks within a period (as in the filters centred at harmonics 15 and 23). As a result, the lower resolvable harmonics dominate the perceived pitch. Such a variation in the clarity of temporal information across frequency accounts well for the findings of Moore *et al.* (1984) described above. (For further details and a more explicit description of the model, see Chapter 5.)

This pattern of outputs across the hypothetical auditory filter bank arises because of the relationship between auditory filter bandwidths and the fundamental frequency of human voices (which determines the harmonic spacing and hence the degree to which harmonics interact in a given filter). If voice fundamentals were in the 10 to 20 Hz range, there would be no filter for which temporal information would be especially clear, whereas if they were in the 2–3 kHz range, temporal information would be clear even in filters centred at 5 kHz.

Taking this approach at face value, we might expect that the fundamental component would always have the highest discriminability, as it is best resolved. In fact, Moore *et al.* found, for complexes composed of the first 7 to 12 harmonics of a fundamental of 200 Hz, that the frequency of the third harmonic was more often discriminated best, although there were individual differences. This is almost certainly due to an interaction between the properties of the initial frequency analysis and those of the temporally-based measuring mechanism that follows.

Some idea about the nature of the temporal process can be gained by examining frequency discrimination performance for sinusoids, which in this model is unaffected by the preliminary auditory filtering. As we noted above, sinusoidal frequency discrimination (in percentage terms) is best in the region 1 to 2 kHz. In Moore and Glasberg's model this can only result from features of the temporal analysis mechanism, which is assumed to be looking at the times between neural impulses generated in an auditory nerve fibre by the appropriately filtered stimulus. Goldstein and Srulovicz (1977) have shown, using a detailed physiological model, that this dependence of performance on frequency can be well accounted for by the statistics of interspike firing times. The performance obtained with sinusoids is thus completely predictable from properties of the temporal analysis, and a consideration of auditory

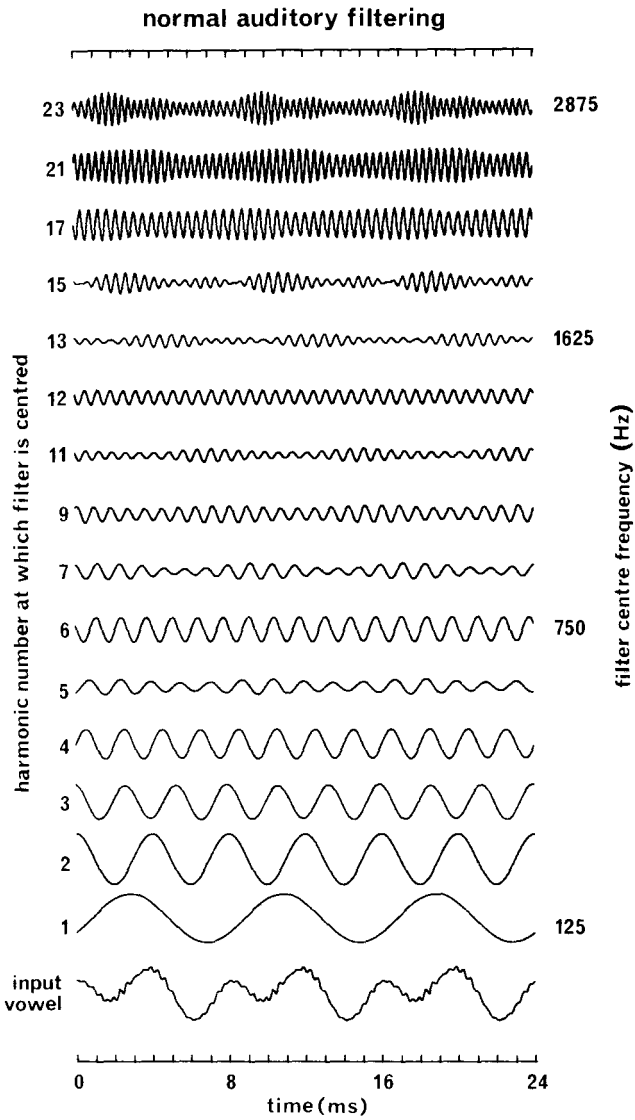


FIG. 7.15 The output of selected channels in a theoretical normal auditory filter bank to a synthetic [i] in which the amplitudes (as given in Fig. 7.14) and phases of the harmonic components are modelled closely on natural speech. At the bottom of the figure is shown the synthetic vowel when synthesized with all 27 of its harmonics. The other waveforms show the outputs from selected hypothetical auditory filters, all of which are centred on a particular harmonic component (indicated at the left of the figure). The centre frequencies of some of these filters are indicated at the right. Filter outputs were calculated by assuming the auditory filters to have an amplitude response given by a rounded exponential, and with the equivalent rectangular bandwidths given by Moore and Glasberg (1983b – also see Chapters 3 and 5). The phase relations used to synthesize the filtered waveforms were the same as those

(continued)

filtering properties is only necessary when more than one sinusoidal component is present.

When harmonic components are equally well resolved (for example, when presented in isolation), performance is best in the mid-frequency region (1–2 kHz). But normally, for fundamentals in the voice pitch range, these higher harmonics suffer interference from adjacent components due to broader filtering. The lowest harmonics, while being best resolved, are less discriminable owing to inherent properties of the temporal measuring mechanism. With temporal acuity increasing, but frequency resolving power decreasing over the first few harmonics, the most discriminable harmonic will probably lie somewhere in the region below 1 kHz, but above 200 Hz, depending upon the fundamental, the precise way in which the auditory filters vary in bandwidth, and the relative amplitudes of adjacent harmonics (see below).

This account explains neatly a rather unexpected result of Klatt's (1973) investigations. In addition to ordinary vowels with fundamentals near 120 Hz, Klatt used vowels which had been high-pass filtered at 500 Hz. This led, contrary to a speculation of Flanagan and Saslow (1958), to reduced jnds in frequency. Clearly, as Klatt hypothesized, the lower components were somehow 'masking' the components that were most important in determining the discriminability of the sounds.

The analysis we have presented is very similar to that seen in Chapter 5 for a train of narrow pulses. Of course, speech sounds, like our [i], have a rather more varied spectral envelope than the equal-component-amplitude sounds usually used by psychoacousticians. This has little effect on the working of the model, however. Variation in harmonic amplitudes may alter the relative importance of particular harmonics in determining the pitch of a speech sound; it will not change the fact that the resolvable harmonics will be the most important, and these will tend to be the lower, if not necessarily the lowest, in frequency (Moore *et al.*, 1985). The primary difference is that if a particular harmonic has a sufficiently large amplitude relative to its neighbours (e.g. by being near a formant peak) it may dominate the response in its frequency region. The temporal information in that frequency region will then be well preserved, and that particular higher harmonic may dominate in determining the perceived pitch.

---

determined for the original synthetic vowel; this is equivalent to filtering with linear-phase filters. Although such an assumption is unrealistic and clearly needs modification, changes in the assumed phase response of the auditory filters will have no effect on our main conclusions (see the text). The peak-to-peak amplitudes of the waveforms have been scaled logarithmically so that the temporal detail in filter outputs of low intensity (e.g. the filter centred at harmonic 13) can still be seen. This is also more psychoacoustically reasonable than a purely linear scaling, given the compressive functions that relate loudness to sound intensity.

Moore *et al.* (1984) demonstrated this possibility with sounds consisting of seven to nine harmonics of a complex of fundamental 200 Hz. Although harmonic components up to the thirteenth were present, at least the first four harmonics were missing. Raising the amplitude of the tenth harmonic not only increased the extent to which its frequency could be discriminated (in an otherwise stable complex), but also decreased the frequency jnd of the complex as a whole. Thus, this rather high harmonic came to dominate the discriminability of the entire complex.

Although this result is suggestive, it is unlikely that harmonics as high as this are typically very important in determining the pitch of speech sounds, at least when heard in quiet. Moore *et al.* intentionally left out the lower harmonics in order that the tenth harmonic could come to dominate. Performance with the lower harmonics included is considerably better than that obtained in the reduced complex, even when the tenth harmonic is as much as 9 dB more intense than the others. So, in circumstances where the lower harmonics are present at a reasonable amplitude (as in speech), they would still be the most important.

On the other hand, it could easily happen that harmonics above the fifth could, if their amplitudes were high enough, become dominant in determining the discriminability of the whole sound. In fact, such a situation might be likely to occur for the vowel we have already analysed. Note in Fig. 7.14 the relative amplitudes of harmonics 5, 6 and 7. The sixth harmonic is some 14 dB more intense than either of its neighbours, most likely as a result of the harmonics immediately above it being attenuated by the vocal tract transfer function (with a low F_1 and high F_2), and the harmonic immediately below it having an unusually low amplitude in the spectrum of the laryngeal source. Whatever the reasons, its increased local amplitude makes it dominate its neighbours, preserving its simple sinusoidal temporal pattern at the output of the auditory filter centred on it (Fig. 7.15). Since this harmonic is effectively resolved from the complex, its discriminability will approach that found in isolation. As discriminability for sinusoids improves with increasing frequency (up to about 1 kHz), this harmonic (at 750 Hz) should be more important in determining the discriminability of the whole complex than the lower, equally well resolved, four harmonics.

Similarly, the analysis of Fig. 7.15 implies that perhaps even the twelfth harmonic, which also gives rise to a near sinusoidal output at 1.5 kHz, could come to dominate discriminability. Care must be exercised in making such claims, however, as the predictions of the model in many cases will be fairly sensitive to specific assumptions about filter shape and bandwidths. Here we have only tried to illustrate its workings in a general way, without worrying too much about the quantitative details.

Although the model outlined by Moore and Glasberg gives, as we have seen, a useful point of view, it does not, on its own, predict many of the

features of normal voice pitch perception. For one thing, it has only been applied to the perception of steady-state pitches, and much work is needed to make it applicable to dynamic contours. Klatt (1973) points out that the results he obtained with simple ramps, which showed a factor of 7 worsening in discriminability when compared with flat contours, already constrain greatly the type of processing it is appropriate to hypothesize. On the other hand, it may be fairly claimed that the main aim of the model is to explore the importance of relatively peripheral (physiologically speaking) factors on the perception of pitch, and this is best done by minimizing the role of more central processes by using steady-state sounds. Elaboration to account for dynamic sounds can surely follow.

In any case, the model is a good framework with which to qualitatively explain many results. Perhaps one of its most important functions is that it can give useful insights into the perception of voice pitch in the hearing impaired, a relatively unexplored area.

v Voice pitch perception in the hearing impaired

There are two primary reasons why the ability of the hearing impaired to hear changes in voice fundamental frequency has been rarely measured. For one thing, the importance of such information, especially in conjunction with lipreading, has only recently come to be appreciated widely (e.g. Fourcin *et al.*, 1979; Risberg, 1974). Its assumed irrelevance is most clearly seen in the standard materials for speech audiometry, with their emphasis on segmental vowel and consonantal contrasts. Secondly, there is the feeling that features related to voice pitch are, on the whole, the most resistant to degradation, whether by noise, filtering or impaired hearing. Although this has been frequently demonstrated for the segmental feature of voicing (see Walden, 1984 for a review), explicit tests of the ability of the hearing impaired to hear changes in voice pitch are fairly uncommon (except in work with cochlear implants, see Schindler and Merzenich, 1985) and, furthermore, require careful interpretation.

Fourcin (1976), in what may be the first explicit test of this type, used a set of synthetic 'oh's' /əʊ/ whose pitch contours formed a continuum which ranged from an extreme rise to an extreme fall. Of six teenage hearing-impaired listeners, three failed to classify the stimuli appropriately as questions (the rising end of the continuum) or statements (the falling end). On the other hand, when asked to label three synthesized vowels ('ee' [i], 'ah' [ɑ] and 'oo' [u]), three performed perfectly, and only one poorly (less than 80% correct). Must we conclude then, at least in this population, that the perception of spectral envelope is superior to that of fundamental frequency? Probably not.

For one reason or another (and perhaps mainly because of its rather weak representation in our system of writing), listeners often have great difficulty

in labelling appropriately utterances that differ in voice pitch contour. Fourcin (1976) notes that the 'perception of pitch change presented more difficulty in general than either of the other pattern forms, *both for the deaf children and for normally hearing subjects*' (italics ours). In our experience, it is not uncommon to find normally hearing adults who are unable to correctly label octave rises and falls, although none (who are literate) have any difficulty in labelling vowels as 'ee' or 'ah' or 'oo'. It may well be that failure in a task of this sort is an expression of a cognitive, not a sensory, limitation.

Experimenters primarily interested in sensory processing have taken two routes out of this difficulty. One that we have used ourselves is to assure, through explicit training (often with a visual display like the Voiscope® (Fourcin *et al.*, 1978; Abberton *et al.*, 1985)) that the listener understands the nature of the task completely. Psychoacousticians, who encounter the same problem with two-interval forced-choice tasks that require a labelling of the direction of pitch movement, often use a 'same–different' task (e.g. Hoekstra, 1979) or three-interval 'oddity' task (e.g. Tyler *et al.*, 1983) instead.

Finally, whether in speech or psychophysical testing, the attribution to a sensory cause of a failure to label fundamental frequency changes correctly is always most convincing when some change in the acoustic parameters enables the listener to perform well. In other words, if listeners always do badly, this may not indicate anything about their basic perceptual abilities. In most of the data we discuss in the rest of this section, listeners show a range of performances which can often be sensibly related to the nature of the acoustic stimuli and the nature of the listeners' impairment.

Before we go on to examine the relevant data, what does Moore and Glasberg's model for pitch perception lead us to expect of the ability of *impaired* listeners to hear changes in voice pitch? Tyler in the previous chapter has shown that most of the hearing impaired suffer from some degradation in frequency selectivity. This may be modelled by simply increasing the bandwidths of the auditory filters (here by a factor of 3) which precede the temporal analysis in Moore and Glasberg's model. If we look at the waveforms at the outputs of these auditory filters (as we did for normal-bandwidth filters in Fig. 7.15) for the same [i] vowel, quite a different picture to the previous one emerges (see Fig. 7.16).

Because most of the filter bandwidths are greater than the inter-harmonic spacing, most channels show a waveform that is, temporally speaking, complex. Only the responses to the two lowest harmonics have the temporally simple patterns that were found for the narrow, low-frequency filters in the simulation of normal frequency selectivity. Even the filter centred on the fundamental shows a nonsinusoidal output, because of the influence of the second harmonic. The output of the filter centred at the sixth harmonic, which might well be the most important in determining the discriminability of the entire sound for normal listeners, is much too complex to play a similar role

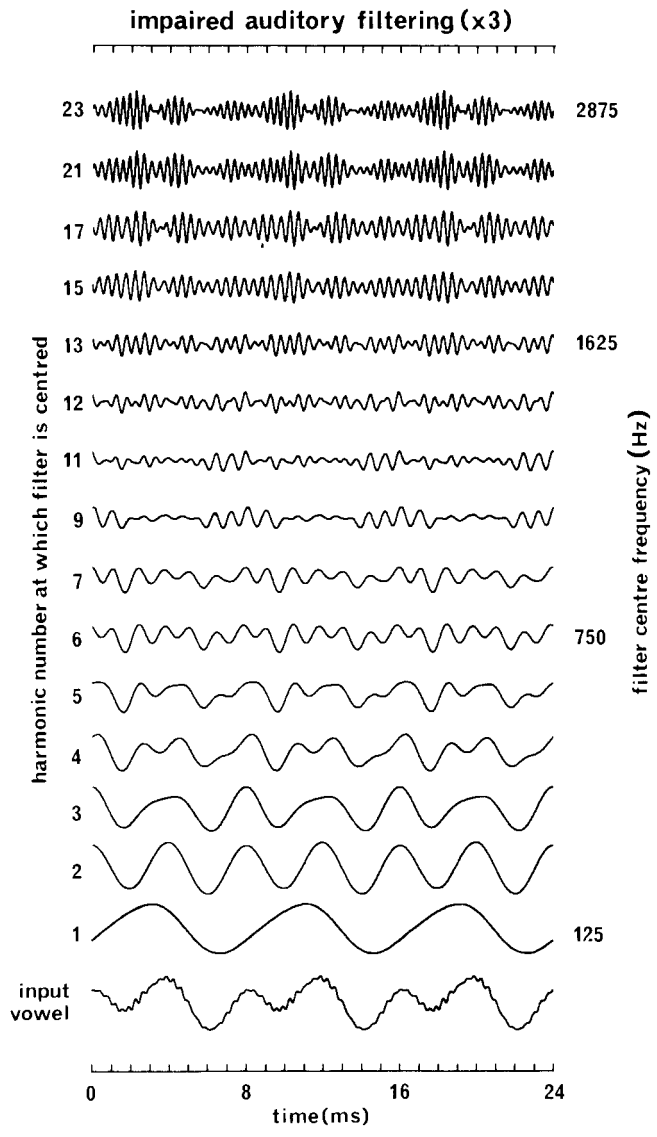


FIG. 7.16 The output of selected channels in a theoretical impaired auditory filter bank to a synthetic [i]. The only difference between this figure, and Fig. 7.15, is that the auditory filters were presumed to have equivalent rectangular bandwidths three times larger than those assumed for Fig. 7.15. Note that the trend in auditory filter bandwidth (increasing with increasing frequency) remained the same.

for the impaired. We should therefore expect a reduced ability to discriminate fundamental frequency changes in complex tones consequent upon reduced frequency selectivity. If auditory filter bandwidths were made yet wider, no channel would contain a temporally simple waveform, and performance would be correspondingly worse. It is important to note that the model allows the extraction of fundamental frequency even from channels where the harmonics interact; it is just that a larger degree of interaction will typically lead to less salient pitches.

Of course, there is more to this model than frequency selectivity, and more to hearing impairment than reduced frequency selectivity. In particular, there is the temporally based processing necessary to measure the time between neural spikes initiated by the filtered waveform. Therefore, a reduced ability to discriminate frequency differences could also result from impaired *temporal* processing, a common occurrence in the hearing-impaired (Moore, 1983). The strength of Moore and Glasberg's model is that it allows a disentangling of these two factors, as the model posits they are independent processes.* Impaired temporal resolution will affect the frequency discrimination of both complex tones and sinusoids, whereas impaired frequency resolution will only affect the frequency discrimination of complex tones.

Some evidence bearing on this issue has already been presented by Moore and Glasberg in Chapter 5 who argued (on the basis of three studies) that there was no relationship between frequency selectivity and sinusoidal frequency discrimination. This result (strongly contrary to place-based accounts of pitch perception) occurs because frequency analysis has no effect on the temporal properties of isolated sinusoids. It is only when multiple sinusoidal components are present that the frequency analysing mechanism is engaged and may make its presence known. Hence, it is much more likely that a correlation will be found between frequency selectivity and frequency discrimination for complex tones. Little such work with impaired listeners has yet been done, a notable exception being the pioneering studies of Hoekstra and Ritsma (1977) and Hoekstra (1979).

*Although it is likely that most of the hearing-impaired suffer deficits in both temporal and frequency resolution, it seems that there is not a close connection between the relative degrees of impairment in each domain. Wightman (1982) for example, has found two impaired listeners with very similar audiograms and psychophysical tuning curves who differ significantly in temporal resolution abilities. Also, Dreschler (1983) has found only weak (although significant) correlations between measures of time and frequency resolution. Other evidence in favour of the relative independence of these two abilities comes from studies (reviewed in Chapter 5) which show nearly normal sinusoidal frequency discrimination in the presence of impaired frequency selectivity (Hoekstra and Ritsma, 1977; Wightman, 1982; Tyler *et al.*, 1983). Of course, this argument makes the assumption that sinusoidal frequency discrimination is primarily, if not exclusively, a temporal process, just as in the model of Moore and Glasberg.

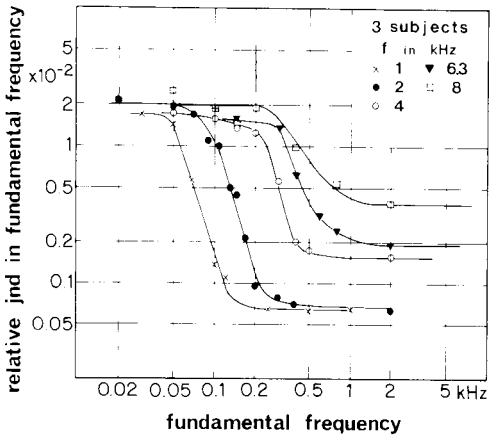


FIG. 7.17 The jnd in fundamental frequency (on the ordinate), as a function of fundamental frequency (on the abscissa) for pulse trains that are bandpass filtered at a variety of centre frequencies, *f*, given in the inset. These results are an average from three normal listeners. The curves have been fitted by eye to the data points. Note how the curves are shifted rightwards as the filter centre frequency is increased. Reproduced with permission and minor modifications from Hoekstra (1979).

They used a rather different approach from the ones we have discussed so far, involving the determination of the jnd in fundamental frequency for 1/3-octave bandpass-filtered pulse trains across a wide range of fundamentals. When frequency jnds are expressed as proportions, normal listeners show a characteristic backward S-shaped curve (Fig. 7.17), as the fundamental frequency of the pulse train is varied from about 20 Hz to the centre frequency of the bandpass filter (1 to 8 kHz in these studies). At this point, the output of the filter is a single sinusoid.

The transition between the plateau of relatively poor performance at low fundamentals, and the relatively good performance at high ones, was attributed by Hoekstra and Ritsma to the properties of auditory filtering. At high fundamentals, the harmonics making up the complex tone are resolved, and performance approaches that found for a pure tone in isolation. At low fundamentals, when the inter-harmonic spacing becomes small relative to the bandwidth of the auditory filters near the centre frequency of the 1/3-octave filter, the harmonics interact, leading to worsened discrimination (just as in the study by Moore *et al.*, 1984). Support for this explanation also comes from the way in which the transition region shifts in frequency with shifts in the centre frequency of the bandpass filter. Just as auditory filters become wider with changes in centre frequency from 1 to 8 kHz, so does the transition region shift upwards in frequency. Clearly, this explanation is completely consonant with the model of Moore and Glasberg we have discussed above.

These same tests were applied to a number of hearing-impaired listeners. Furthermore, in an attempt to explain the individual results, measures of frequency selectivity (psychophysical tuning curves—PTCs—discussed in Chapters 3 and 4) were also taken. On the whole, the results from the two tests concur. From the point of view of Moore and Glasberg's model, if a patient has normal temporal resolution but impaired (yet not absent) frequency resolution, s/he should also show a backwards S-curve, but with the transition region shifted towards higher fundamentals (i.e. greater harmonic spacing). This was, in fact, a frequent finding. However, the position of the transition region is liable to give an overestimate of frequency selectivity once a patient's auditory filters get much wider than 1/3 octave. As long as the temporal analysing mechanisms are relatively intact, there will still be a backwards S-curve, no matter what auditory filtering is present, because the stimulating waveform still changes from a harmonic complex waveform to a simple sinusoid as fundamental frequency is increased. Now though, the changes in the waveform, and hence the position of the transition region, will be determined by the 1/3-octave external filter. In a sense, the electrical filter substitutes for the auditory filter. This does not affect the results from normal listeners because their auditory filters are somewhat narrower than 1/3 octave, at least at the frequencies used here.

Other listeners showed much more deviant results, to the extent of the curve losing its bend altogether, mostly because performance dramatically worsened at high fundamentals (where the acoustic stimulus became a sinusoid). Hoekstra (1979) hypothesized that this was due to a complete loss of frequency selectivity, an explanation which assumes that sinusoidal frequency discrimination is mediated primarily by a place mechanism. Interestingly, such completely flat curves were almost always obtained for a filter centre frequency of 4 kHz, where it is likely, on both behavioural and physiological evidence, that place codes *are* more important. At frequencies of 4 kHz and above, we might well expect sinusoidal frequency discrimination and frequency selectivity to be related, even if such a correlation does not hold at lower frequencies.

It seems likely that most of Hoekstra's results can be explained on the basis of a model like Moore and Glasberg's, as long as careful attention is paid to two points: (1) the relative size of the bandwidths of the auditory and external filters, and (2) the relative importance of temporal processes at low (1 and 2 kHz) and high (4 kHz) filter frequencies.

There are also two difficulties, one of which is essentially contradictory to Moore and Glasberg's model. The more minor problem relates to the way in which hearing impairment affects temporal analysis. As was noted above, when impaired listeners gave flatter backwards S-shaped curves than normal, it was primarily due to performance worsening for high fundamentals of the pulse train, that is, for sinusoids. In fact, the average frequency jnd for

sinusoids in the impaired listeners was significantly worse than the average for untrained normal listeners, while for low fundamentals the average discriminability in the impaired listeners was only slightly worse than the normal average. This led Hoekstra (1979) to conclude that 'for many hearing impaired listeners the discrimination of temporal features such as the waveform envelope repetition rate is fairly normal.' As mentioned above, this explanation is consistent with Moore and Glasberg's model only at 4 kHz, but flatter than normal curves were also found at lower frequencies. Here, poor performance with sinusoids should be a reflection mostly of a deficit in temporal processing ability. It may well be, though, that temporal analysis in the impaired is damaged in such a way that the coding of low stimulation rates is relatively unaffected. This is consistent with the physiological findings of Woolf *et al.* (1981) who investigated the degree of phase-locking in the auditory nerve and ventral cochlear nucleus (VCN) of kanamycin-deafened chinchillas. In both sets of neurones (and especially in the VCN), it seems that phase-locking was much less seriously degraded at low stimulation frequencies (<0.75 kHz) than at high.

The more serious difficulty for Moore and Glasberg's model is Hoekstra's finding that normal backward S-shaped frequency discrimination curves (whose transition region should be related to the auditory filter bandwidth at that filter frequency) can sometimes be obtained in the presence of degraded auditory filtering. It is interesting to note, however, that in these cases, so-called W-shaped PTCs were obtained, which are difficult to interpret (see Chapter 6 for a further discussion of this). It may well be that frequency selectivity is not nearly so impaired as the W-shaped PTC indicates.

We have discussed this set of studies so extensively because they were the first to look in detail at the perception of complex tones by the hearing impaired. As they are aimed at characterizing basic psychophysical processes, however, the stimuli are rather different from those typical of voiced speech. This makes it difficult to apply the results directly to the difficulties encountered in speech perception by impaired listeners, especially as Hoekstra *et al.* did not investigate the frequency region likely to be most important in determining the pitch of voiced speech sounds, i.e. that below 1 kHz. Similar studies with sounds more closely related to speech are needed.

Some preliminary work along these lines has been initiated in our laboratory, spurred initially by the desire to determine optimal ways to present voice pitch information to the profoundly hearing impaired. Normal listeners, as was seen above, have more acute discrimination with complex tones because they are able to combine information across harmonics that are relatively well resolved by the auditory filters. If frequency selectivity is minimal, it may well be that sinusoids are better discriminated than complex tones, simply because they present a simpler waveform to the temporal processor.

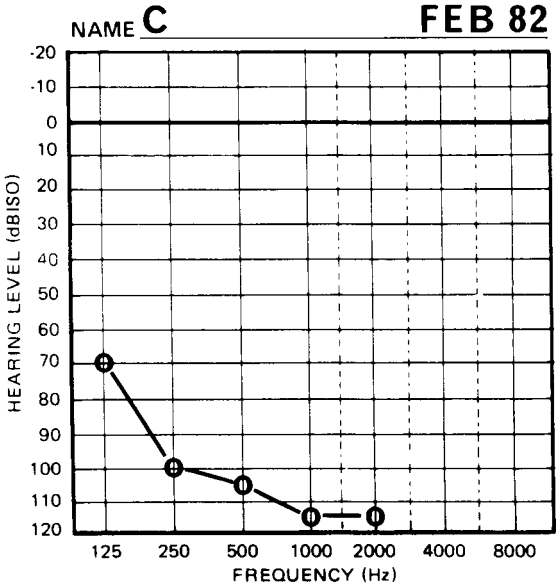


FIG. 7.18 A pure-tone audiogram for patient C. Thresholds were unobtainable at 120 dB HL for 4 and 8 kHz.

This was first demonstrated by Rosen and Fourcin (1983) in extensive investigations on one profoundly deafened patient, C, whose audiogram is shown in Fig. 7.18.

C, unlike normal listeners, showed significantly better frequency discrimination performance for sinusoids than for trains of narrow pulses at fundamental frequencies of 100 and 240 Hz. Stock and Rosen (1986) obtained the same outcome in another patient with a similar (although less severe) loss, for a fundamental frequency of 125 Hz. This result is completely analogous to that of Moore *et al.* (1984) who determined the discriminability of components in otherwise fixed harmonic complexes for normal listeners. Recall that although low harmonics (essentially resolved by auditory filtering) had jnds similar to those found for sinusoids in isolation, higher unresolved harmonics (due to wider auditory filters) were poorly discriminated in comparison to isolated sinusoids. Clearly, it is the latter condition that is relevant for the hearing-impaired, whose auditory filters are probably wide compared to the inter-harmonic spacing.

Unfortunately, there is no independent measure of frequency selectivity for either of these patients, mostly because of the problems of getting any masking at all when thresholds are already so high (see Chapter 6 for a further discussion of this point). However, with two other listeners having rather more moderate losses, Stock and Rosen were able to measure selectivity using

a two-point version of the notched-noise technique (with normalized notch widths of 0 and 0.4 at a noise spectrum level of 60 dB SPL/Hz; see Chapter 3 for further information on this technique). Both listeners had reduced selectivity at two or three of the three frequencies used (125, 250 and 500 Hz—designed to span the normal adult voice pitch range), and did not always exhibit lower jnds for pulse trains than for sinusoids.

Particularly striking were the different patterns exhibited by the two listeners. M had a sloping loss (10 dB HL at 125 Hz to 80 at 1 kHz), normal selectivity at 125 Hz, slightly impaired selectivity at 250 Hz, and possibly none at 500 Hz. He showed a more or less equivalent performance with sinusoids and pulse trains at all three frequencies. W, on the other hand, had a flat loss (about 40 to 50 dB HL), practically no selectivity at 125 Hz (where sinusoids were slightly better discriminated than pulse trains), and demonstrable, but still impaired, selectivity at 500 Hz (where pulse trains were discriminated significantly better). Such results may be expected if harmonics in the upper frequency region (0.5–1 kHz) must be sufficiently resolved for improved performance with multi-component sounds.

Furthermore, although W's jnds for sinusoids at 125 Hz were essentially equal (within 15%) to those of the relatively unpractised normal listener in the study, they were rather larger for pulse trains (by some 140%). This is consistent with the demonstrated sharper filtering of the normal listener and the supposition that W's temporal processing at 125 Hz (which needs to be explicitly investigated) is relatively unimpaired. Although there are still difficulties in interpreting many of the details of the data, the results lend strong support to the idea that the degree of frequency selectivity may be a much more important factor in the discrimination of fundamental frequency differences for complex tones, than for sinusoids.

Such phenomena are not confined to the relatively unnatural, static, psychoacoustical stimuli discussed so far. Rosen and Fourcin (1983) also tested C's ability to label (as rising or falling) six different synthesized fundamental frequency contours on a continuum whose endpoints were schematized versions of natural voice pitch contours (seen in the lower left-hand corner of Fig. 7.19). These were presented in three possible 1-octave frequency ranges (80–160 Hz, 130–260 Hz and 200–400 Hz), and in three different acoustic forms; as sinusoids, pulse trains and on a synthetic vowel. Figure 7.19 shows the resulting categorization curves.

Using the slope of each curve as a measure of discriminability (more steeply sloping curves indicating higher discriminability), the following points may be noted:

1 C is totally unable to label the high fundamental frequency range (200–400 Hz) stimuli when they are presented as vowels. As the /ɑ/-vowel used has a high first formant, it might be thought that this is due

to the fundamental being relatively low in amplitude. However, when pulse trains are used (with the lower harmonics of essentially equal amplitude), labelling accuracy is still at chance.

2 Performance with sinusoids is always better than that obtained with speech, but only in the highest frequency range is this difference statistically significant.

The superiority of sinusoids over complex tones is also seen when C tries to identify the intonation pattern of natural sentence-length utterances spoken as questions (with a final rising intonation contour) or statements (with a final falling contour). Two speakers, a male and a female, were used, and again, the fundamental frequency contours could be presented in a number of acoustic forms. In addition to three conditions similar to those above (natural speech, pulse trains triggered by vocal fold closures, and frequency-modulated sinusoids synchronized to the instantaneous voice pitch of the speaker), there was yet another condition. The stimuli were also frequency-modulated sinusoids, but here the instantaneous frequency presented to the listener was 50 Hz lower than the actual voice fundamental frequency of the speaker. This not only lowers the contour, it also stretches it in ratio terms (Fourcin *et al.*, 1984).

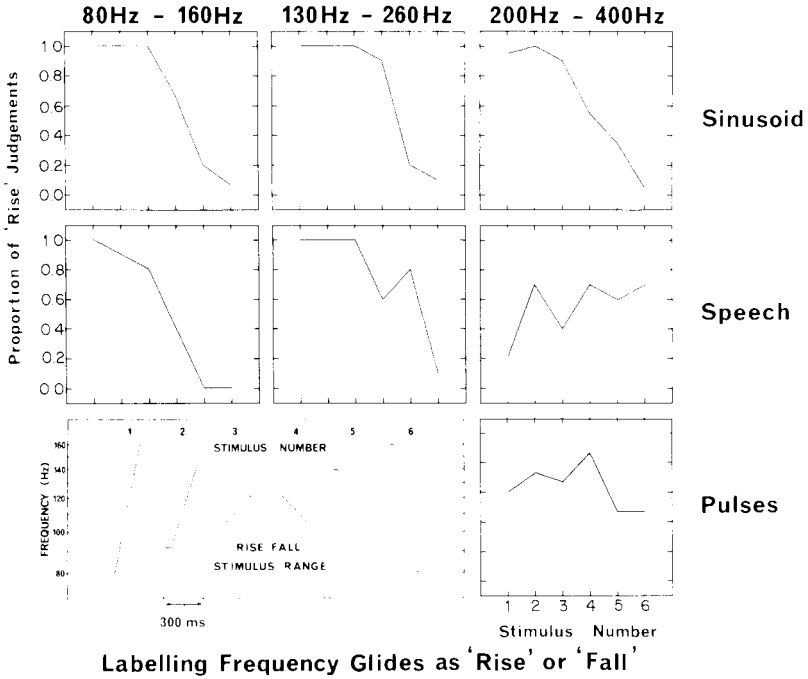


FIG. 7.19 Categorization functions obtained from patient C for frequency glides drawn from a continuum spanning an octave rise and fall, for three different types of sound (sinusoid, vowel or pulse train) presented in three different frequency ranges. Pulse trains were not used in the two lower frequency ranges. For further details, see the text and Rosen and Fourcin (1983).

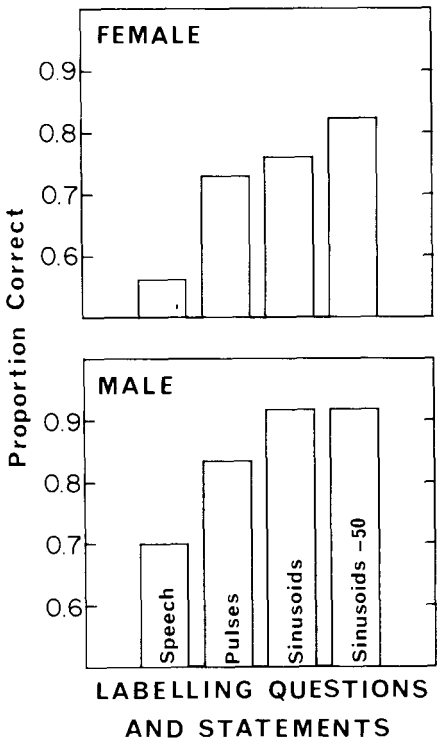


FIG. 7.20 Proportion correct obtained by patient C in a natural speech question/statement identification task in which the speech was processed in several ways. For further details, see the text and Rosen and Fourcin (1983)

For both male and female speakers, labelling accuracy is worst for speech and best for sinusoids, as seen in Fig. 7.20. For the female speaker only (with a higher voice fundamental), sinusoids mapped down in frequency are superior to unmapped sinusoids. Performance with pulse trains falls in between. That the labelling of pulse trains is better than the labelling of natural speech probably has little to do with frequency selectivity, but more to do with the fact that pulse trains have neither the amplitude nor the spectral fluctuations present in speech. (The clinical implications of these findings will be discussed in Section 4D.)

There is one further way in which reduced frequency selectivity can affect the perception of voice pitch. This is by introducing a degree of phase sensitivity much greater than that found in normal listeners. This possibility was first demonstrated in a small experiment by Hoekstra and Ritsma (1977) and Hoekstra (1979), albeit for stimuli only remotely related to speech. They tested the ability of a number of hearing-impaired listeners to hear changes in modulation frequency of either 100% sinusoidally amplitude-modulated (SAM) or quasi-frequency-modulated (QFM) complex tones.

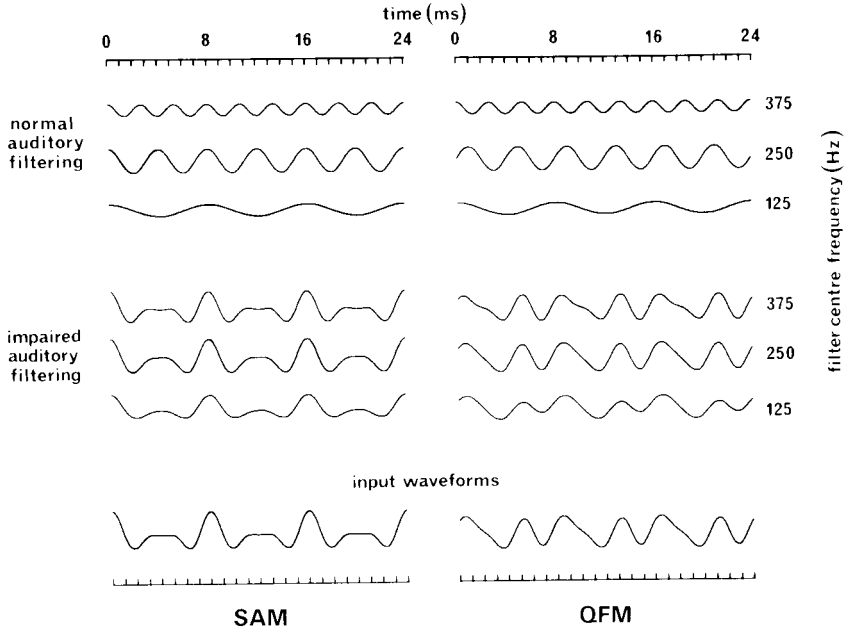


FIG. 7.21 The outputs of hypothetical normal and impaired auditory filter banks to two three-component harmonic complexes which differ only in the relative phase of their central component. The auditory filters are centred at the frequencies of the harmonic components. Normal auditory filtering uses the rounded exponential model and bandwidths given by Moore and Glasberg (1983b) as in Fig. 7.15, while impaired auditory filtering assumes the same trend of bandwidth with frequency, but with absolute values ten times larger than in the normal case.

Such SAM and QFM complexes have identical power spectra (two components, equally spaced from, and 6 dB down in amplitude from the central carrier), but differ in their phase spectra, and hence their time waveforms. For the SAM complex, all components are in phase (sine phase in Hoekstra's work) while for the QFM complex, the central component is 90° out of phase. Thus the SAM stimulus has a much more 'peaky' amplitude envelope than the QFM stimulus. Hoekstra (1979) reported that of five hearing-impaired listeners tested, three were significantly better in discriminating changes in modulating frequencies near 200 Hz for the SAM complex than for the QFM complex, at a carrier frequency of 2 kHz. Normal listeners (and the other two impaired listeners) showed no difference between the two conditions. We have also shown (Rosen, 1986b) in other impaired listeners (including patient C), that relative phase can be a strong determinant of jnds in fundamental frequency for a rather more speech-related sound (the first three harmonics of a complex tone with a fundamental near 250 Hz).

Results like this are easy to understand from the point of view of Moore and Glasberg's model. Figure 7.21 shows, for a set of normal and widened (here by a factor of 10) auditory filters, the outputs of filters centred at each harmonic to SAM and QFM stimuli which are the first three harmonics of a complex tone at a fundamental of 125 Hz.

Note that for a 'normal' auditory filter bank, the outputs are identical for the two phase conditions (at least with regard to interpeak periods). Thus it is not surprising that phase has little or no effect. For the impaired case, however, there are large differences at the filter outputs to SAM and QFM complexes. The SAM complex has clearly defined peaks at the fundamental in all channels. The filtered versions of the QFM complex (and clearly the filters are so wide relative to the harmonic spacing that they exert only a minor influence on the waveforms) have a much flatter envelope, with the intervals between the major peaks having two durations, neither related simply to the fundamental. Furthermore, these vary across the filters shown, with frequencies corresponding to 217–207 Hz and 295–315 Hz, with increasing auditory filter centre frequency. We might well expect the pitch extracted from such waveforms to be ambiguous, thus leading to large jnds.

Of course, real auditory systems will introduce their own phase shifts into the signals, a transformation not included in the model we have presented here. The phase changes introduced by impaired systems are likely to differ significantly from those introduced by normal ones, and to differ widely among the impaired. It may well turn out, then, that harmonic complexes having the most salient pitch are *not* those that have their components in phase at the tympanic membrane, and that the particular phase relationship leading to the most salient pitch will vary among listeners.

It is interesting to note that the models of Moore and Glasberg (Chapter 5), van Noorden (1982), and Srulovicz and Goldstein (1983), all predict phase insensitivity for resolved components because they all work from interspike firing times in the auditory nerve. Seneff (1984), on the other hand, has proposed a physiologically based model for voice pitch extraction that operates on the sum of the auditory filter outputs (after compression and rectification). This model cannot be correct because it is highly phase sensitive in a way which does not depend on the sharpness of auditory filtering.

That phase relations among harmonic components may be important in determining the salience of voice pitch changes also has important implications for understanding the speech perception of impaired listeners in free-field conditions. Close to a speaker's mouth, voiced speech waveforms typically have a single major peak per fundamental period, associated with the fast closure of the vocal folds. These rather peaky waveforms may, like SAM stimuli, be heard with a strong pitch even by the hearing-impaired.

In a reverberant sound field, however, the phases of the harmonics become completely random (Plomp and Steeneken, 1973) often leading to relatively flat-envelope waveforms, more similar to QFM complexes. The pitches of these may not be heard very distinctly at all. In this way, perceptual abilities measured with headphones, or free-field in very 'dead' rooms, may seriously overestimate the perceptual capabilities of the impaired in more everyday circumstances.

One final point, discussed again later (Section 3Bxii), needs to be made. Since it is generally understood that monaural phase sensitivity is constrained by auditory filtering (Mathes and Miller, 1947; Goldstein, 1967) it might be expected that a loss in frequency selectivity would lead to more acute phase sensitivity in the hearing impaired. (Such an outcome would be consistent with the hypothesis that the temporal information available to the central auditory system is restricted to the interspike firing times on each auditory nerve fibre—see Fig. 7.21.) This has, in fact, been recently demonstrated by Rosen (1984, 1986a,b), who showed that at least some hearing-impaired patients (with demonstrable loss in frequency selectivity) could discriminate SAM from QFM stimuli better than any normal listener at the same intensity levels. Changes in phase relation are often reported as subjective changes in pitch, yet another demonstration of the importance of temporal processes in determining the pitch of complex tones.

B *Static spectral patterns*

Although we have already noted that no speech sound remains constant for long, there are important sound classes for which essential aspects of their acoustic and perceptual properties are usefully represented by a quasi-steady-state description: vowels, fricatives and nasals. Even for essentially dynamic sounds like plosive stops, there are acoustic cues to their identity that are essentially static.

i Importance in speech contrasts

Given a particular type of excitation (either voiced, voiceless or mixed), the primary determinant of the shape of the spectrum will be the shape of the vocal tract. Hence, it is not surprising that the phonetic features cued by spectrum shape are those associated with particular articulatory configurations. Classical phoneticians, whose descriptive terms rely heavily on (at least putative) details of production, term this feature *place of articulation* (at least when dealing with consonants) which refers to the point in the vocal tract where the principal articulators come closest together, or touch. Hence, the English sounds /m/, /n/ and /ŋ/ (as in Ki*m*, ki*n* and ki*ng*), all of which are voiced nasals, differ on this dimension (being bilabial, alveolar and velar, respectively) as do the voiceless fricatives /f/, /θ/, /s/, /ʃ/

and /h/ (as in *f*ie, *th*igh, *s*igh, *sh*y and *h*igh, being labiodental, dental, alveolar, palato-alveolar and glottal, respectively). Similarly, the voiced fricatives /v/, /ð/, /z/ and /ʒ/ (as in *v*at, *th*at, *z*ap and plea*s*ure), the counterpoints to the first four voiceless fricatives, can also be ordered in this way along the place dimension.

Well-defined articulatory positions which lead to differences in spectral shape can also be useful in distinguishing what are otherwise considered fast changing sounds, the plosives. The transient bursts typically found at their onset have a spectrum that essentially reflects the articulatory configuration at release. That such information is actually used by listeners has been demonstrated by Pols and Schouten (1982), who showed that, on average, the isolated bursts led to better identification performance than did the following transitions on their own, although the latter also contributed importantly to plosive identity.

Vowels can also be distinguished from one another on the basis of steady-state articulatory and, hence, acoustic configurations. In fact, vowel quality differences are usually the first thing to come to mind when discussing the importance of static spectral features for speech perception, a tendency also indicated by the degree to which work on vowel contrasts dominates this area.

In phonetic terminology, vowels are not normally considered to vary along a single place-of-articulation dimension, but to be characterized by their values along a number of dimensions, all, however, related to features of vocal tract shape. Figure 7.22 shows the relationship between the formant patterns of vowels and the traditional articulatory descriptors of high/low (or equivalently close/open) and front/back which were thought to indicate the position of the highest point of the tongue. It is now apparent from X-ray studies, which show significant differences among speakers, that these descriptors are more a reflection of auditory facts than physical ones, but in that sense they do adequately describe the perceptual relationships between vowels [see Joos (1948, pp. 54–56) and Ladefoged (1975) for a further discussion of this issue].

Static spectral patterns can also cue some distinctions related to the phonetic features of *manner of articulation* and voicing. Contrasts between vowels, nasals and nasalized vowels, although they can involve steady-state differences in the shape of the vocal tract, are usually labelled as manner differences by phoneticians. Also, since the voiced source has its greatest energy at low frequencies, while voiceless excitation tends to be weighted in energy towards higher frequencies, steady-state spectral patterns can aid the voiced/voiceless distinction, at least in continuant sounds like fricatives and nasals.

Finally, just as for fundamental frequency, spectral features can convey much paralinguistic information, for example, the sex, age, accent and identity of the speaker. In general, formant frequencies are inversely related to vocal tract length, so, on average, formants are higher in frequency for children than for women, and for women than for men.

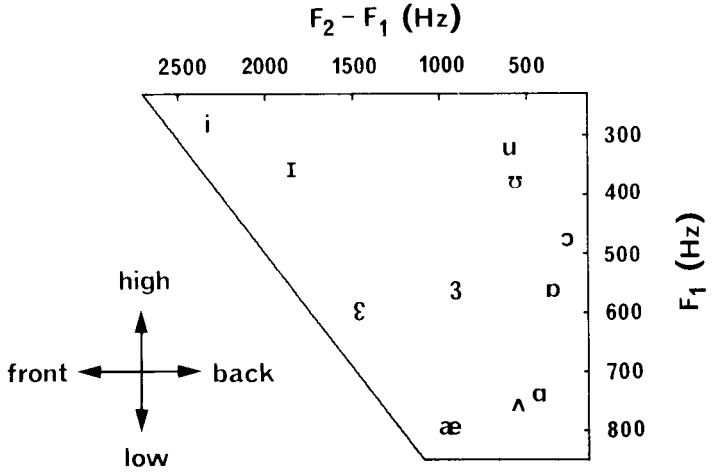


FIG. 7.22 A vowel quadrilateral for the relatively pure vowels of British 'Received Pronunciation' (RP). The formant values reported in an unpublished thesis by J. C. Wells (reproduced in Gimson, 1962) were used to determine the appropriate place of each vowel in the quadrilateral. Note how the classical phonetic features of high/low (equivalently close/open) and front/back correlate with the frequency of F_1 and the difference between F_1 and F_2, respectively. As will be seen by comparing this figure to Ladefoged's (1975) similar representation of American English vowels (although they too will vary from region to region), there are many vowels in American and British English which have at least roughly similar qualities: /i/ as in '*h*e*e*d', /ɪ/ as in '*h*i*d*', /ɛ/ as in *h*e*a*d', /æ/ as in '*h*a*d*', /ʌ/ as in '*H*u*dd*', /ɑ/ as in *h*a*rd*' (RP) or '*h*o*d*' (American), /ɔ/ as in '*h*a*w*e*d*', /ʊ/ as in '*h*oo*d*' and /u/ as in '*wh*o'*d*'. /ɜ/ occurs in RP pronunciations of words like '*her*' and '*bird*', and is also like the hesitation noise ('uh') used in English on both sides of the Atlantic. /ɒ/, the vowel in RP '*h*o*d*', does not have a close equivalent in General American English, but a similar vowel is used in and around Boston (just as in RP, e.g. '*h*o*d*' or '*b*o*dy*'). See Wells (1982) for detailed discussions of the various vowel systems used in English throughout the world.

In summary, quasi-steady-state spectral pattern features can make it possible to: distinguish sounds within the classes of fricatives, nasals, plosives and vowels; separate nasals from vowels from nasalized vowels; assist the voiced/voiceless contrast; and provide paralinguistic information about the speaker.

ii Major acoustic features

Let us deal with the simplest case first, non-nasalized vowels. As discussed in Section 2B, their spectra can be modelled as the product of a harmonic source spectrum and a vocal tract transfer function which has only simple resonances, producing peaks in the output spectrum which are known as formants. Only the three lowest in frequency seem crucial for vowel quality judgements, and their positions vary widely for different vowels. Fant (1956) claims, in agreement with the measurements of Peterson and Barney (1952), that for an average male speaker, the range of F_1 is 150–850 Hz, of F_2, 500–2500 Hz, and of F_3, 1700–3500 Hz. Women and children's formants are higher, with the maximum third formant frequency a little above 4000 Hz.

Formant bandwidths for men vary between about 40 and 200 Hz, with *Q*s (the ratio of centre frequency to bandwidth of the resonance) from about 7.5 to 35 (Dunn, 1961; Fant, 1956). Fujimura and Lindqvist (1971) give a range of about 30–120 Hz but this is for a closed glottis condition which, with decreased damping, should give narrower bandwidths (see footnote, p. 383). The formants of women tend to be wider than those of men (Fujimara and Lindqvist, 1971). There is some controversy regarding the way in which the bandwidth of a formant is related to its frequency. Dunn's results show bandwidth to increase nearly uniformly with formant frequency while Fujimara and Lindqvist claim that the bandwidth of F_1 increases as its frequency decreases below 400 or 500 Hz. In any case, it seems fair to say that, on average, formant bandwidth increases with formant number.

When sounds are nasalized, increased damping in the nasal cavity causes a general broadening of the formant bandwidths. Also, antiresonances typically lead to the appearance of dips in the spectrum. Fujimura (1962) showed that the main antiresonance increases in frequency (from about 1 to 3 kHz) as the place of articulation for the English nasal consonants moves from front to back (from /m/ through /n/ to /ŋ/). Furthermore, he claims two other features (a low F_1, well separated from the other formants, and a high density of upper formants in the frequency domain) which lead, in all three nasals, to a relatively even distribution of energy in the frequency region between about 800 and 2300 Hz. This property, in combination with the existence of the low F_1, is hypothesized to be an important one in distinguishing the steady-state parts of nasals (known as the nasal murmur) from other vowel-like sounds.

A consideration of fricatives and plosive bursts introduces new acoustic features. Firstly, much of the spectral detail differentiating them occurs in rather higher frequency regions than does the information for voiced sounds. Heinz and Stevens (1961) found that a sample of the English sounds /f/, /s/ and /ʃ/ could be well matched by white noise passing through a system with two resonances and an antiresonance (as long as some low-frequency modifications were made to the noise for /f/ and /ʃ/). The resonance locations varied from 2.2 to 12.2 kHz, depending upon the fricative, with the antiresonance in the range from 2.3 to 6.8 kHz. The farther forward the place of articulation, the higher the frequency of the resonances, since they arise from the cavity in front of the constriction, shorter cavities having higher resonances. Secondly—a factor which may be crucial for listeners—the source for fricatives is inherently random. Hence, the spectrum of voiceless sounds will vary from moment to moment, and this may add a further complication for the perceiver attempting to extract the spectral pattern. Plosive bursts are characterized by being rather short in time, and this may also pose extra difficulties for a listener.

The type of description of acoustic properties we have just given is by far the most common one, although it may be a poor one for a consideration of auditory processing of speech. Studies of the physical properties of speech sounds are normally related to an articulatory/acoustic model, and hence are reported in terms of resonances and antiresonances. From the point of view of audition, what is likely to be more useful is, for example, a characterization in terms of the height and position of peaks in the spectrum and the depth of the valleys between (even though these properties are, of course, completely predictable from the resonance/antiresonance description). A first approach in this direction has been made by Plomp (1983) who prefers to talk in terms of 'modulation' of the spectrum, as opposed to the specification of formant frequencies.

iii The discrimination performance of normal listeners with speech-like sounds

The first modern study which attempted to quantify the sensitivity of listeners to changes in spectral patterns at least somewhat similar to speech was that of Stevens (1952).* He passed single narrow pulses through a parallel RLC circuit (which can be represented in the frequency domain as a single resonance) to generate damped sinusoids. Through appropriate manipulation of the circuit parameters, the centre frequency and bandwidth of this resonance could be varied. The jnd in centre frequency for a range of centre frequencies (0.2 to 5.0 kHz) and bandwidths (about 16 to 160 Hz) was measured. Although Stevens did not discuss his results in relation to speech, we can think of it as an experiment in single-formant centre-frequency discrimination (where the formant exists for a single voice-pitch period!) at least for the bandwidths that would be sensible for speech (above 40 Hz).

Two important findings may be noted. Firstly, discrimination performance at all centre frequencies improved as the resonance bandwidth decreased. Secondly, the absolute jnd (in Hz) at a particular bandwidth tended to increase with increasing frequency, for centre frequencies above 500 Hz. There was, however, a tendency, especially for the wider bandwidth resonances, for performance to worsen as centre frequency decreased from 500 to 200 Hz.

These two main findings have been replicated by a number of workers using stimuli that are, in differing degrees, more closely related to speech. Horst (1982) and Ritsma *et al.* (1967, 1968) used longer duration (225 ms and 500 ms, respectively) bandpass filtered pulse-trains resembling single formants near 1 and 2 kHz, with fundamental frequencies in the range 40 to 140 Hz. Discrimination performance improved with increasing spectral

*Interestingly, Joos (1948, p. 83) reports formant frequency jnds of a constant 6% (except for vowels in the low-back region) from a pilot experiment using two-formant vowels. This accords well with the values found in later, more extensive, studies.

slope (i.e. decreasing bandwidth) for slopes up to about 140–160 dB/octave, roughly equivalent to the narrowest formant bandwidths in real speech. Gagné and Zurek (1980) measured the just-discriminable centre-frequency change in a single formant in the F_1 range (300 to 800 Hz) and found performance to be better with decreasing formant bandwidth (with formant Qs in the range 1 to 36).

All studies since Stevens' show the absolute jnd to increase with increasing centre frequency (Horst, 1982; Mermelstein, 1978), three even in the region where Stevens found the opposite effect (Flanagan, 1955; Gagné and Zurek, 1980; Pickett and Martony, 1970; the first changed F_1 or F_2 in a four-formant vowel, while the latter two used a single formant). Flanagan, and Gagné and Zurek only used F_1s down to 300 Hz, where the upturn Stevens found was small. Pickett and Martony, however, made measurements from 205 to 825 Hz with single formant stimuli, and their jnds are smallest at 205 Hz. All in all, the bulk of the data (as well as the theoretical models we discuss below), suggest that absolute jnds always increase with increasing formant frequency. Perhaps the anomalous result of Stevens arises from his use of a single-pulse stimulus.

Although nearly all reports agree about the general way in which discriminability varies with centre frequency and bandwidth, there are fairly large discrepancies in the absolute values of the jnds reported. These seem due to a combination of differences in the acoustic structure of the stimuli, the nature of the task, the method used to calculate the jnd, the listeners used and the extent of their listening experience. For relatively unpractised listeners, jnds are on the order of 2 to 5% across the range 0.2 to 2.0 kHz. The best performance is reported, as is usual, in the psychoacoustically oriented literature. Horst (1982), for example, used stimuli with triangular spectra centred at 1.2 and 2.0 kHz and found jnds better than 0.5% for bandwidths appropriate for speech (Qs of about 15–20 according to Dunn (1961), which Horst states are equivalent to his spectral slopes of about 90–120 dB per octave). Both Flanagan (1955) and Pickett and Martony (1970) found rather poorer performance near 1 kHz, with jnds between 2 and 5%, but both studies used formant bandwidths that were rather large (about 145 Hz). Even at these bandwidths, though (a Q of about 7, equivalent to a spectral slope of about 40), Horst's jnds were rather smaller, being about 1%. A complicating factor is that there was an extra cue in Horst's stimuli—the fundamental was always a constant proportion of the resonance centre frequency. Hence it, as well as the spectral envelope, changed. However, the jnds were still about 1% when fundamental frequency was held constant.

This last issue raises a problem that dogs all such research, and makes comparison across studies difficult. If one wants to measure 'pure' formant-frequency jnds, clearly the fundamental frequency should be invariant across the two stimuli to be compared. If, however, the fundamental is held constant,

then for high Q formants, the levels of the harmonics can change dramatically, altering the overall level of the sound for small changes in formant centre frequency. This introduces an extra cue for discrimination. For this reason, Horst (1982) usually maintained a constant frequency ratio between the centre frequencies of his 'formants' and his fundamental (i.e. as the formant frequency increases, so does the fundamental, so that the harmonics always fall at the same place relative to the spectral envelope). Of course, this too can make an extra cue available—the change in fundamental. One sensible (and speech-like) way out of this *impasse* would be to use sweeping fundamental frequency contours in place of the more common monotones. This would keep the overall level of sound constant with changing formant frequency, even at narrow bandwidths, and furthermore, would more realistically represent the task faced by listeners in extracting spectral pattern information from speech.

Results of this type have been obtained by Gagné and Zurek (1980). They compared performance in discriminating F_1 frequency changes for fundamental-frequency contours which were varying (a 50 Hz rise and fall) or static. At Q values low for speech (1 and 2.8, i.e. wide formant bandwidths), the form of the pitch contour did not affect discrimination performance. For more reasonable Q-values of 10 and 36, however (i.e. narrower formant bandwidths), the jnds for changing pitch contours were consistently larger than those obtained for a monotone. It seems likely that these larger jnds represent the 'true' jnd for formant frequency discrimination, uncontaminated by artifactual loudness variation.

The data for aperiodically-excited formant-type stimuli are much more sparse. Ritsma *et al.* (1967, 1968) found discrimination to improve as the spectral slopes of bands of noise near 2 kHz were made steeper. Gabrielsson *et al.* (1975) determined the just-noticeable difference in the centre frequency of relatively wide bands of noise (Qs of 0.4, 0.7 and 1.4, meant to approximate the bandwidths of voiceless fricatives) at frequencies of 0.25, 0.5, 1.0 and 2.0 kHz. They found, just as for periodically-excited formants, that the jnd increased with increasing frequency and decreased with increasing Q.

The extent to which the random fluctuations in aperiodically-excited formants impair performance relative to that found for periodically-excited ones has been investigated even less. Although it seems likely that moment-to-moment variations in spectrum would indeed impair discriminability, the fact that the spectral envelope is continuously defined for aperiodic sounds, instead of at discrete points as with periodic sounds, may mitigate some of the effect of randomness. Unfortunately, the available data are contradictory. Ritsma *et al.* (1967) found no difference in the jnd for resonances of the same Q excited by periodic pulse trains at 140 Hz and by noise. In 1968, however, they found that periodic sounds led to superior performance.

Finally, just as we noted for fundamental frequency patterns, the jnds measured under the ideal conditions used in experiments like these (especially with highly trained listeners) are likely to be considerably smaller than those needed for speech contrasts. Unfortunately, few or no studies have tried to determine the degree of sensitivity that *is* necessary, but some rough estimates can be made from measurements of speech production. Wells (cited in Gimson, 1962), for instance, gives formant frequencies (used to construct Fig. 7.22) measured from 11 relatively pure British English vowels which constitute a fairly rich vowel system. Each vowel differs by at least 12% in F_1 or F_2 from all others, but this is likely to underestimate the necessary acuity, as the utterances were in citation form, and hence more acoustically distinct than they would be in everyday speech.

One of the control procedures in Peterson and Barney's (1952) vowel study is also of interest in this regard. All the words acoustically analysed were uttered twice by each speaker, and Peterson and Barney compared the formant values obtained each time. Presumably, the speakers were attempting to produce exactly the same sound, and so any variability that did exist would not be auditorily significant. Peterson and Barney only show their analyses for one instance of 90 examined (their Fig. 7.7), but there, the mean difference between the pairs of utterances for 28 speakers was 17 Hz for a mean F_1 of 310 Hz, a change of about 6%. It thus seems reasonable to suppose that only formant frequency differences of 6% or more will play an important role in speech contrasts.

iv Excitation-pattern models of spectral feature discrimination

Although there are many details lacking, at least the major findings discussed above can be explained via a set of ideas developed most thoroughly by Zwicker (1970) and his colleagues. In this view, discrimination performance is mediated primarily through the excitation patterns of the sounds, a concept that has been discussed thoroughly in Chapters 3, 4 and 5. In short, an excitation pattern is meant to represent the degree of activity (or excitation) evoked by a particular sound at some unspecified level in the auditory system, as a function of a variable related to frequency (usually the Bark or ERB; see Chapter 5). Excitation patterns can also be thought of as a representation of the amount of activity present at the output of the auditory filter bank, across the set of varying centre frequencies. Hence, they are a direct reflection of the degree of auditory frequency selectivity: for a sound of restricted bandwidth, the narrower the bandwidths of the auditory filters, the narrower the excitation pattern, and the steeper its slopes.

According to Zwicker, listeners can discriminate two sounds when their excitation patterns differ by some criterion amount (usually around 1 dB) at some point on the frequency-related scale. More specifically, in modelling

the discrimination of sounds which differ only in centre frequency, the excitation patterns change most where their slope is greatest. So, the steeper the slope of the excitation patterns, the smaller the jnd in centre frequency will be. As has already been seen in Chapter 5, for sinusoids in normal listeners the greatest change occurs on the low-frequency side of the excitation pattern.

In order to apply this model to what we know about formant-frequency discrimination, it is necessary to determine how the bandwidth and centre-frequency of formants influence the slopes of their excitation patterns. Let us first examine the increase in jnd found with increases in centre frequency, which is predictable from the fact that auditory filter slopes (in dB/Hz) decrease with increasing centre frequency. For a formant of a given bandwidth, the higher its centre frequency, the shallower will be the slopes of its excitation pattern. So, a larger shift in centre frequency will be necessary at the higher frequencies to cause the 1-dB change in level between the two patterns necessary for discrimination.

An alternative way of looking at this uses Zwicker's (1970) assertion that when plotted on a critical-band-rate scale (in Bark units), auditory frequency selectivity is constant across frequency (see Chapter 5). Thus, excitation patterns for sounds which differ solely in centre frequency are simply lateral translations of one another. Since the critical band scale is roughly logarithmic (at least over the greatest part of the range of most interest in speech (Zwicker and Terhardt, 1980; Moore and Glasberg, 1983b)), the jnd for formant frequency should grow approximately in proportion with centre frequency (i.e. Weber's Law applies). Hence, the absolute size of the jnd will grow with frequency.

The dependence of discriminability on formant bandwidth arises from the fact that when the physical spectrum has shallower slopes than those of the relevant auditory filters, the slope of the excitation pattern is determined by the slope of the spectrum of the incoming sound. As the maximal slope of a single formant is approximately proportional to its Q-value (i.e. narrower bandwidths lead to steeper slopes; Horst, 1982), better discrimination performance will occur for narrower bandwidth formants. When, however, the spectral slopes of the input sound become steep relative to the slopes of the auditory filters, discrimination performance is predicted to level off with further sharpening of the formants. In this case, the slopes of the excitation patterns depend almost solely on the properties of the auditory filters and performance should then equal that found for a pure tone. Both these predictions have been borne out in Horst's (1982) study.

v Psychoacoustical measurements of the excitation patterns of speech sounds

Although it was once thought that the simultaneous masking pattern of a sound could serve as a good indication of its excitation pattern, there are

now serious doubts about the adequacy of this measure. In particular, as has been thoroughly discussed in Chapter 4, the excitation patterns inferred from paradigms in which masker and probe are nonsimultaneously presented typically retain much more spectral detail than those obtained in tasks where masker and probe are simultaneous. As excitation patterns have proved to be a very fruitful way to approach many problems in psychoacoustics, it is only natural that there have been efforts to assess the accuracy of current techniques for determining them. This is especially so for sounds with a relatively complex spectral shape, like steady-state vowels, for which the importance of nonlinear phenomena (like suppression) is not well understood.

Moore and Glasberg (1983a), for instance, measured the masking patterns for two synthetic vowels in both simultaneous and nonsimultaneous (forward) masking, and found that the vowel formants were much more clearly represented in forward masking (see Fig. 4.23). Sometimes the difference between the formant peaks and valleys in the masking patterns was actually larger than it was in the physical stimulus, a finding previously reported by Houtgast (1974), who was the first to apply nonsimultaneous techniques to vowels (using the pulsation-threshold method).

Because of its important implications (some of which we examine below), much more work of this kind needs to be done. For example, as the average formant spacing is constant across frequency, and the bandwidths of both formants and auditory filters increase with frequency, we should expect the resolution of formants to decrease with increasing frequency. Resolution of formants should also be better for narrower formant bandwidths and larger formant spacing. There may well be details in the spectrum that are not perceivable by human listeners, and hence need not be preserved in the outputs of speech synthesizers (especially for fricatives in the region where auditory filtering is crudest). Few data are available to address any of these issues.

vi Psychoacoustical models of spectral pattern perception

We now turn our attention to attempts at characterizing subjective responses to spectral envelope variations on a more global level than those involved simply with explaining the discrimination and peripheral representation of spectral features. Such work usually goes under the name of 'timbre research', but care is necessary with this term. Spectral envelope is only one of five acoustic correlates of timbre that Schouten (1968) lists, but both because of its importance in speech, and its interaction with frequency selectivity, it is the only one we will discuss here. (He also includes: the range between tonal and noiselike character; the time envelope in terms of rise, duration and decay; the gradual changes both in spectral envelope and fundamental frequency; and, the prefix, an onset of a sound quite dissimilar to the ensuing lasting vibration.)

As long as we are dealing with sounds generated by the same source of excitation (either voiced, voiceless or mixed), there is a fairly large body of work in psychoacoustics which, although not specifically aimed at understanding the perception of speech sounds, is applicable to them. In the approach promulgated most extensively by Plomp (1976 – Chapter 6) and his co-workers, timbre (in our limited sense) is argued to be a direct correlate of the physical amplitude spectrum of the sound, processed through the auditory filters. This implies that for the periodic stimuli typically used, the phase relationships among the individual component harmonics do not affect the timbre of a sound. Although known not to be strictly true (Licklider, 1957; Schroeder, 1959), it is a reasonable approximation when speech sounds are processed by normal listeners.

Plomp and Steeneken (1969) quantitatively estimated the importance of phase for a set of ten-component harmonic complexes with fundamental frequencies in the range 146 to 585 Hz. Initial studies showed that component phase relationships affected timbre maximally when one of the sounds had all components in either sine or cosine phase, and the other had components alternating between sine and cosine phase. A set of sounds with harmonics in these two phase relationships was synthesized. The slope of the spectral envelope was also varied. Listeners compared all the stimuli in triads, and, for each triad, were asked to pick the pair of stimuli which were most similar, and the pair which were least similar. These judgements were used to obtain a measure of the subjective distances between the stimuli, and so to compare the effects of phase and spectral tilt. Although there was a clear effect of phase on timbre, it was small compared to the effect of varying the amplitude spectrum. This has led Plomp and his colleagues in their later work to concentrate on a sound's amplitude spectrum only.

In this approach, the auditory filter bank is modelled as a set of 1/3-octave filters. The 'timbre' of any particular sound is represented as a vector whose coordinates indicate the level (in dB) of the output from each of these filters. The difference in timbre between two sounds can then be predicted with a distance metric (usually Euclidean) calculated from their respective vectors. The adequacy of the theory is assessed by comparing the subjective differences (calculated from the triadic comparisons) with the physically derived ones.

Alternatively, a psychological 'map' obtained from multidimensional scaling of the results of triadic comparisons is compared with a physical 'map' obtained from a principal-components analysis of the vectors obtained from the 1/3-octave filter bank. (Principal-components analysis is a way of reducing the number of dimensions necessary to describe a set of stimuli by finding, via a geometrical rotation, a new set of dimensions which accounts for the greatest variance in the data. Typically, the 15–18 dimensions of the original measurements (number of 1/3-octave filters) is reduced to two to four dimensions.) Good agreement between these two 'maps', and/or high

correlations between the subjective and physical differences, has been obtained for two sets of tones; one derived from nine different musical instruments, and one from ten stops of a pipe organ (Plomp, 1970, 1976).

vii Applying psychoacoustical timbre models to speech sounds

Essentially the same set of procedures has been applied to vowels, with great success. The first complete study of this type was reported by Pols *et al.* (1969). Steady-state vowels were constructed by excising one cycle from the vocalic part of 11 different CVC words uttered by a single male speaker, and modifying all 11 to have identical duration. For presentation to listeners, these single cycles were played repetitively from a computer for 405 ms, after being adjusted for equal loudness. In this way, the 11 stimuli judged were of equal loudness, duration and fundamental frequency, and differed only in their spectral envelope. As in the previous studies, the results of triadic comparisons underwent multidimensional scaling to construct a psychological 'map', and the outputs of a 1/3-octave filter bank underwent principal-components analysis to construct a physical 'map'. These two 'maps' were very similar, as can be seen in Fig. 7.23. Furthermore, as had already been found in an earlier study (Plomp *et al.*, 1967), there was a close relationship between the 'map' derived from the two most important dimensions extracted by the principal components analysis and a plot of the first *vs* second formant frequency.

This pattern of findings has been replicated by Klein *et al.* (1970) in a study using rather more natural stimuli (100-ms segments of 12 vowels from each of 50 male speakers). The physical analyses were done as before, but here the psychological map was constructed from the confusions which arose in an identification test.

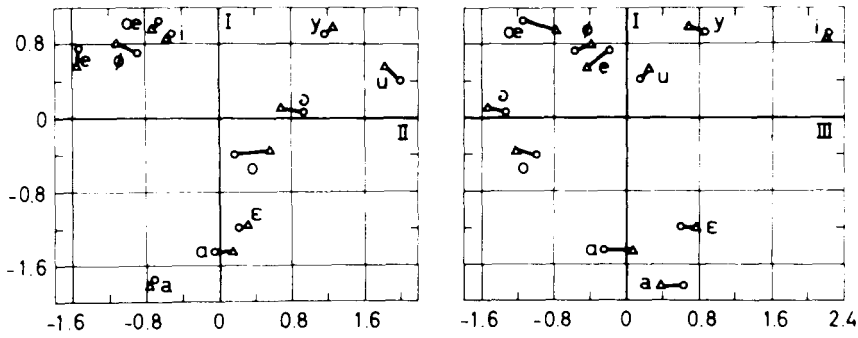
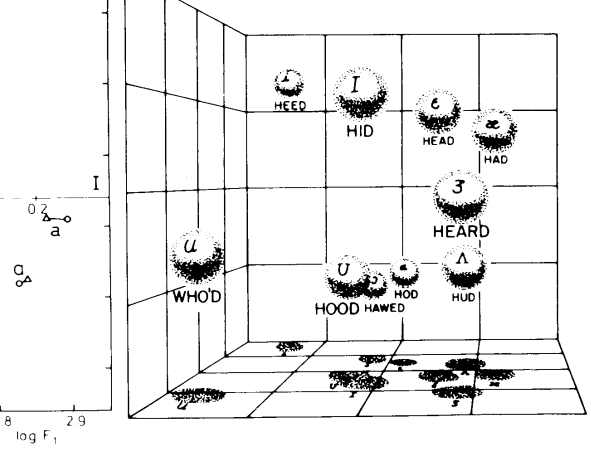


FIG. 7.23 Positions of points in the optimal I-II and I-III planes when the six-dimensional 'map' obtained by principal-components analysis of the acoustic structure of the given vowels (circles) was matched with the three-dimensional perceptual map obtained by multidimensional scaling of triadic comparisons (triangles). For each vowel, the resulting values from each type of scaling are connected together with a line. Reproduced with permission from Pols *et al.* (1969).


HEED
HID
HEAD
HAD
HEARD
WHO'D
HOOD
HAWED
HOD
HUD
I
a
log F1

Although even these stimuli are fairly unnatural, the Dutch workers have argued for the validity of their approach, stressing the close correspondence between results from speech and nonspeech sounds, and the importance of overall spectral shape in each. Supporting evidence comes from the original multidimensional scaling performed by Shepard (1972) on the confusions reported by Peterson and Barney (1952) in their classic study. Here the sounds were ten /h-vowel-d/ utterances recorded from a total of 76 speakers, including men, women and children. Seventy listeners were asked to identify each of the utterances. Shepard was able to use the errors from this experiment to obtain the psychological 'map' shown on the right side of Fig. 7.24. On the left are some representative data from Pols *et al.* (1973), showing the relationship between the processed 1/3-octave filter data and the formant frequencies of each vowel. Clearly, allowing for the different vowel systems used in American English and Dutch, the first two dimensions of the Shepard analysis are very similar to those of Pols *et al.* This similarity arises because the results of perceptual scaling, factor-type analyses of spectral shapes and a formant analysis give essentially the same results.

viii Is vowel quality determined by spectral pattern shape or the position of formants?

In one sense, the filter-bank based approach goes against a long tradition of acoustic phonetic research in which it is considered that the crucial attributes of vowels, productively, acoustically and perceptually, are the positions of the formants. The Dutch group have tended to take a fairly agnostic position in this controversy, noting the relative advantages and disadvantages of each approach. For example, Pols *et al.* (1973) have contrasted the closer relation between formant specifications and vocal tract shapes with the much greater ease of appropriately weighting a 1/3-octave

FIG. 7.24 *(opposite)* A comparison of three different representations of the relationships among vowel sounds. On the left is a comparison of two methods for scaling the physical differences between 12 vowels spoken by 50 male speakers, one based on formants, and one based on 1/3-octave spectra. The scaling of the 1/3-octave filter-bank analyses was done in a different way than the principal-components analysis used in Fig. 7.23. First, the original cluster of points in 18-dimensional space (i.e. from a bank of 18 filters) was factor analysed to determine the three-dimensional subspace which best accounted for the variance in 18 dimensions. Then, the plane through that space that gave the best identification performance across the vowels was determined (the so-called 'maximally discriminating plane'). The triangles represent the best match of the mean vowel points in this 'maximally discriminating plane' to the configuration obtained by plotting log F_2 *vs* log F_1 (circles). Part of the reason why this figure looks so different from Fig. 7.23 is that solutions from factor-type analyses and multidimensional scaling can be freely rotated and flipped, as these transformations preserve interpoint distances. On the right is a multidimensional scaling analysis performed by Shepard (1972) of the confusions made by listeners in an identification experiment. The original three-dimensional solution was rotated to match, as well as was possible, the frequency values of the first three formants measured from the original stimuli. For further details, see the text. The figure on the left is reproduced with permission from Pols *et al.* (1973), and that on the right reproduced with permission from Shepard (1972).

spectrum than of extracting formants.* They go on to assert that whether or not 'formant extractors' operate in vowel perception is a question that can only be decided by further experimentation.

One way to explore this issue is through the measurement of what is known as F_2'. It has long been known that only two formants are necessary to synthesize a vowel which sounds fairly satisfactory (Joos, 1948), although the two lowest formant frequencies from the complete vowel are not necessarily those that give the best match of vowel quality in the reduced, two-formant vowel (Delattre *et al.*, 1952). When the first formant of the

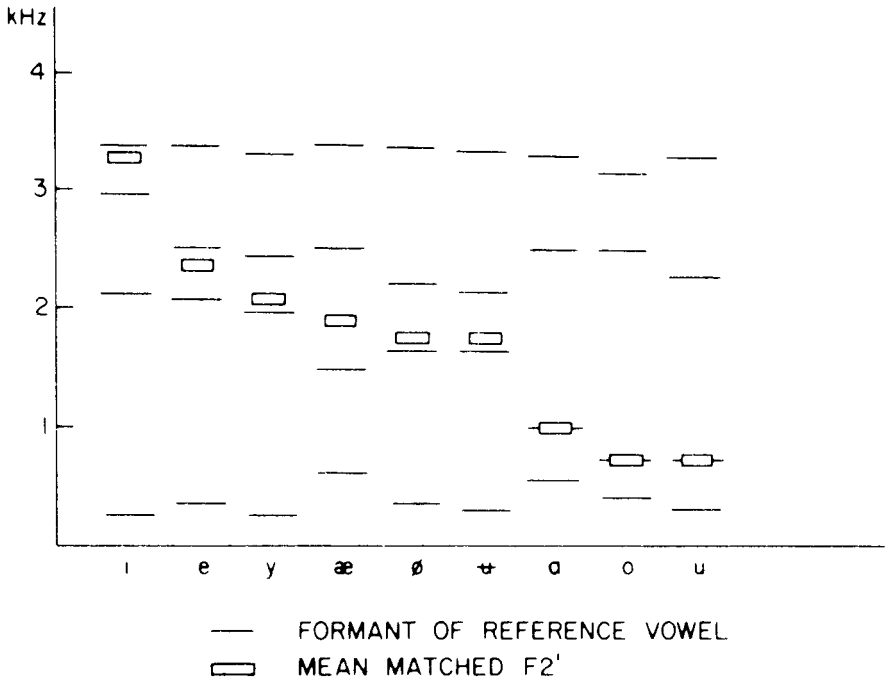


FIG. 7.25 The results of an experiment in which listeners were asked to adjust the frequency of the second formant in a two-formant vowel to best match the quality of a four-formant vowel which had the same value of F_1. For further details, see the text. Reproduced with permission from Carlson *et al.* (1970).

*Many workers in the field have noted the difficulties in estimating formant frequencies from physical spectra, a task which often necessitates *a priori* knowledge of where the formants *should* be for a particular vowel – as seen already in Fig. 7.5, where the two lower formants of [ɑ] were not separately resolvable at the higher fundamental frequency. These difficulties tend to be especially great for female speakers. van Nierop *et al.* (1973), in a study of 12 vowels from each of 25 adult female speakers, report difficulties in determining the frequency of a formant in about 40% of the 300 vowel segments.

complete and reduced vowel are at the same frequency, F_2' is the frequency of the second formant in the reduced vowel that gives the best match in quality to the complete (usually four-formant) vowel. Figure 7.25 shows the results of a matching experiment in which listeners were allowed to adjust F_2' in order to make the best match to a complete vowel (Carlson *et al.*, 1970).

When F_2 and F_1 are close together, with F_3 fairly far away, as for the three vowels on the right side of the figure, then F_2' is equal to F_2. Once F_2 gets a little further from F_1, however, F_2' is always higher in frequency than F_2. In this case, when F_3 is closer to F_2 than to F_4, F_2' is somewhere between F_2 and F_3 (although /æ/ may be an exception). Finally, when F_3 comes close to F_4, (as for /i/) then F_2' falls between them.

In order to avoid the criticism that listeners were operating purely in a psychoacoustical mode with no relation to speech perception, Carlson *et al.* also performed an identification test. Here, listeners were asked to label (using appropriate Swedish orthography) a large set of two-formant vowels in which the first and second formant frequencies were varied. In this way it was possible to map regions in the F_1–F_2' plane which corresponded to each vowel category. There was good agreement between these regions and the results of the adjustment experiment.

At first sight, it may seem that data like these are incompatible with theories which posit formant extraction or, equivalently, the detection of spectral peaks, as the foundation of vowel perception. This judgement may be premature until the effects of limited auditory frequency selectivity are taken into account. If the auditory system always resolved all the formants (which we know it does not), then F_2' experiments would provide strong evidence against formant extraction models. On the other hand, since the internal auditory spectrum is a smoothed version of the physical spectrum (as in the models of Plomp (1970) and Zwicker (1970)), groups of formants may be represented by single peaks in the excitation pattern. A peak in the excitation pattern might occur which is not present in the physical spectrum.

Carlson *et al.* (1970) have evaluated this idea with a model based on the basilar membrane tuning found by von Békésy (the tuning is now understood to be much sharper than this; see Chapter 1). Picking amplitude peaks from the hypothesized distribution of activity across the membrane gave some promising results, although with significant difficulties. They then combined this model with a set of counters which measured the number of zero-crossings in the waveform at the output of each of the 120 filter channels which made up the model basilar membrane. A histogram, constructed by counting the number of channels responding at a particular frequency, always gave two or three isolated peaks. The two highest peaks obtained from a four-formant vowel corresponded in frequency with the F_1 and F_2' appropriate for that vowel.

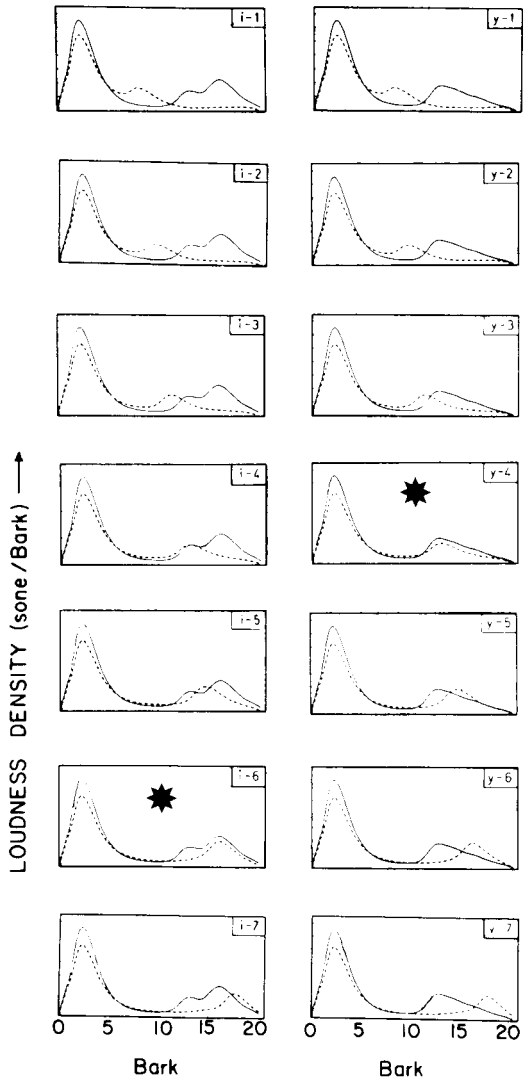


FIG. 7.26 Calculated auditory spectra from a theoretical model incorporating a frequency scale transformation (Hz to Barks), critical-band filtering and a transformation from physical intensity units (dB) to perceptual ones (sones). Each panel shows the spectra corresponding to two vowels. In the column at left, the solid lines (constant from top to bottom) show the auditory spectrum corresponding to a four-formant synthetic Swedish [i]. The dashed lines are the spectra resulting from a series of two-formant vowels in which the value of the first formant coincided with that of the first formant in [i], and in which the second formant frequency varied from 1 kHz (at top) to 4 kHz (at bottom) in 1/3-octave steps. In the column at right, the solid lines represent the auditory spectrum of a four-formant synthetic Swedish [y], while the dashed lines show the auditory spectra calculated from the same two-formant vowels as in the column at left. (The two full-formant vowels had the same F_1.) The four-formant vowels were paired with each of the two-formant vowels (but not with each other), and listeners were asked to rate their similarity. The 'starred'

(continued)

Further evidence of the possible adequacy of formant-extracting mechanisms is found in the studies of Bladon and Lindblom (1981), though this was not their intention. They have essentially extended the timbre model of the Dutch group by incorporating a theory of auditory processing rather more elaborate than 1/3-octave filtering. In their model, both frequency selectivity and the transformation from a physical energy unit (dB) to a psychological one (sone) are meant to reflect auditory properties more accurately (see Section 2 of Chapter 5). The model has been used to predict subjectively judged differences between vowels as a function of their spectral differences, and also in accounting for the position of F_2' for two of the vowels used by Carlson *et al.* (1970).

In these latter experiments, a complete vowel was paired with a set of two-formant vowels in which the position of the second formant was varied. Listeners were asked to rate (on a five-point scale) the similarity of the two vowels. Figure 7.26 shows the 'auditory spectra' calculated by Bladon and Lindblom for each of the comparison stimuli along with the 'auditory spectra' calculated for the standards. The two-formant stimuli which, according to their model, should give the best match are indicated by stars, and indeed these were the pairs judged most similar. It is interesting to note, however, that identical results would have been obtained with the rule 'Match the highest peak in the auditory spectrum (outside of the fixed F_1)' as were obtained with the rule they used: 'Minimize the root-mean-square difference between the two curves'.

Whether spectrum-matching and peak-picking rules would always give equivalent answers remains to be seen. If they do not, it may turn out that variation in the metric used to determine distances may make them more similar. In particular, both Plomp (1970, 1975) and Bladon and Lindblom have used a metric of the form:

$$D_{i,j} = \left[\sum_{n=1}^{m} |EX_{i,n} - EX_{j,n}|^p \right]^{1/p}$$

where $D_{i,j}$ is the distance between two excitation patterns, EX_i and EX_j, each of which arises from a bank of m auditory filters, and $EX_{i,n}$ is the excitation from the nth filter channel arising from stimulus i. (If continuous variables are used, as, at least formally, they are in Bladon and Lindblom's work, the summation sign becomes an integral taken over the audible frequency range.)

---

panels show, for each full-formant vowel, which two-formant vowel was judged as most similar in quality. Note that this occurs when the F_2 peak in the two-formant auditory spectrum is at the same position in frequency as the major high-frequency peak in the full-formant auditory spectrum. Reproduced with permission and minor modifications from Bladon and Lindblom (1981).

When $p = 2$, the standard Euclidean distance arises (root-mean-square), and when $p = 1$, the metric measures the differences in area between the two patterns. For vowel sounds, both Plomp, and Bladon and Lindblom investigated the effect of varying p from less than 1 to about 5: p values around 2 were said to account best for the data, but larger values than this led to predictions that were either just as good as those for $p = 2$, or not significantly worse. Furthermore, although Bladon and Lindblom argue that there is a nonlinear relationship between obtained and predicted auditory distances, they use the correlation coefficient (a measure of linear correspondence) to determine the appropriate value of p. All this is to say that p values greater than 2 might be considered. The utility of this arises from the fact that, in the limit, as p goes to infinity the metric above measures the single greatest difference between the two patterns (which is the same criterion used in Zwicker's excitation pattern theory of discrimination). It seems likely that a criterion for F_2' matching that states: 'Minimize the maximum discrepancy between the two patterns' will lead to results very similar to one that requires peaks to be matched. Hence, it may well be that the spectral-pattern and the peak-picking approach can be subsumed into the same general model.

One particularly important issue in this controversy is the degree of auditory filtering actually present. We have already discussed the discrepancy between excitation patterns determined using simultaneous and nonsimultaneous masking techniques. According to the studies of Houtgast (1974) and Moore and Glasberg (1983a), the degree of detail present in the auditory representation of vowels is considerably greater than that found after 1/3-octave filtering, or the smoothing used by Bladon and Lindblom. It is therefore crucial to know how sensitive the two types of model are to variations in assumed frequency selectivity. It is our guess that spectrum shape models will vary less in their predictions with changes in the degree of auditory frequency selectivity than formant-picking models (although Carlson *et al.*'s combination of place and time processes may make that model less sensitive to variations in selectivity).

There are many other investigations that can help to clarify this issue. Thorough reviews have recently appeared by Chistovich and her colleagues, who have been the strongest proponents of some sort of formant extraction, albeit followed by a mechanism for integration of peaks within 3.5 Barks (Chistovich, 1985; Chistovich *et al.*, 1979). Whatever the outcome, it is clear that the specification of frequency selectivity must play a crucial role.

ix Phonetic versus psychoacoustical processing

Before we leave this topic, it might be best to consider a rather more fundamental issue: To what extent can *any* primarily auditory model account for judgements of speech, whether based on formants, overall spectral shape,

or both? Carlson and Granström (1979) present some data that could be interpreted as giving the answer 'very little'. They synthesized a set of 66 variations of a vowel similar to /æ/, in which 13 properties such as formant frequencies, bandwidths, the phase relation among harmonic components and the overall spectral tilt were varied (Carlson *et al.*, 1979). These were paired with the reference /æ/ and presented to listeners for judgements of dissimilarity. Two types of instructions were given. One set of listeners was told to take into account *any* dissimilarity between the vowels, while the other set was told to 'disregard changes associated with harshness, speaker identity, or transmission channel' and 'rate only changes that influence vowel identity'.

The result was that the overall dissimilarity rating given to each of the 13 acoustic manipulations differed greatly with the instructions given. For the 'psychoacoustical' instructions, phase manipulations had the greatest effect while variations in F_1 and F_2 simultaneously were in fifth place. F_2 variations on their own came ninth. With the 'phonetic' instructions, the four combinations of formant frequency manipulations came top of the list, with phase manipulations in fifth place. Given the fairly arbitrary choice of the degree of change on each dimension, and the incommensurability across dimensions, we should not concern ourselves too much with the absolute positions of the factors, but rather with the way they change. How can any psychoacoustical model explain this?

One clue for a way out of this dilemma may be found in the already mentioned study of phase effects in timbre by Plomp and Steeneken (1969). In addition to harmonic complexes with fixed spectral tilts, they also used three vowels, similar to each other in quality, from the study of Pols *et al.* (1969). Each vowel was synthesized in three different versions: (1) with the phase relationships found in the single period of natural speech, (2) with all harmonics in sine phase, and (3) with alternating sine and cosine phase across the harmonics. In a set of previous studies, the phase change between the two latter stimuli had been found to lead to the greatest change in timbre. All nine stimuli were presented for triadic comparisons, and the results scaled to yield a psychological 'map' of the subjective relation between the sounds. A three-dimensional solution (reproduced in Fig. 7.27) gave a very good fit to the observed data. Two of the dimensions were found to relate to vowel quality. On these two dimensions were found three tightly-packed clusters of points, each far from the others, and each containing only a particular vowel, irrespective of phase condition. The third dimension related purely to phase differences, with the sine and natural phase stimuli essentially indistinguishable, and the alternating phase stimuli far away. There seemed to be no interaction between the two factors, these having essentially orthogonal effects on perceived similarity.

A similar conclusion can be drawn from another study of Plomp and Steeneken, in which overall amplitude, spectral tilt and phase relations were

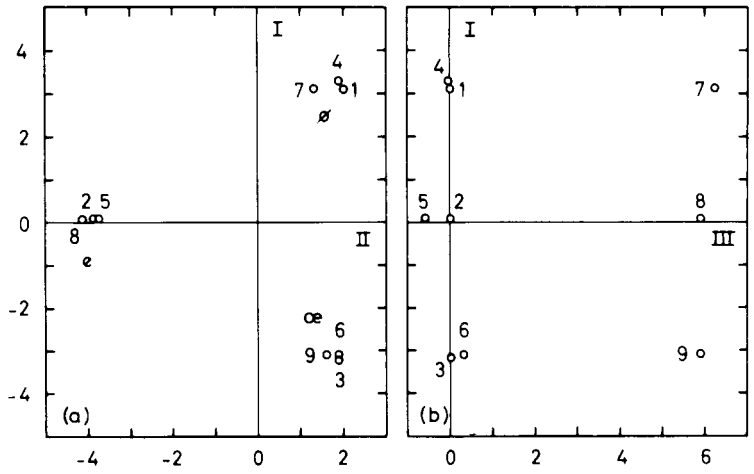


FIG. 7.27 The results of a multidimensional scaling of the judgements from triadic comparisons of nine vowel-like stimuli. All the sounds were purely periodic. Stimuli 1, 2 and 3 were obtained by repetitive playing of a single period extracted from the natural Dutch vowels [ø], [e] and [œ] respectively (after adjustment to make them of equal loudness and fundamental frequency), all of which are similar in quality. Stimuli 4, 5 and 6 had the same amplitude spectra as stimuli 1, 2 and 3, respectively, but their Fourier components were added together in sine phase. Stimuli 7, 8 and 9 again had the same amplitude spectra as stimuli 1, 2 and 3, respectively, but now the Fourier components were added in alternating sine and cosine phase. Thus, the stimuli fell into three groups, in each of which the three sounds had the same vowel amplitude spectrum, and differed only in phase. This figure shows the projections of the I-II and II-III planes of the three-dimensional scaling solution. The I-II plane corresponds to the differences in amplitude spectrum, while dimension III corresponds to the differences in phase spectrum. Note how sine-phase synthesis leads to sounds which are auditorily essentially indistinguishable from those which have the phase relationships of natural speech. Reproduced with permission from Plomp and Steeneken (1969).

varied for nonspeech sounds. Again, three orthogonal dimensions were found to give a good fit to the obtained perceptual data, one for each of the three variables.

Plomp and Steeneken's listeners were instructed to judge the vowel stimuli in essentially a 'psychoacoustical' mode, but we would expect them to give different judgements, as Carlson and Granström's listeners did, if they had been instructed to respond 'phonetically'. This seems especially likely as the effects of spectral shape were orthogonal to those of phase. Such changes in judgement, however, need not require the perceptual space to alter. It may simply be that, depending on the task, listeners vary the importance they attach to differences on the various dimensions. In Carlson and Granström's case, they (sensibly enough) weigh the dimensions pertaining to phonetic identity more heavily when they are instructed to.

This is not a new idea in multidimensional scaling. It has long been recognized that individual perceivers may differentially weigh perceptual dimensions of stimuli, even though the underlying perceptual space is similar.

Algorithms have been developed that allow the extraction of both the dimensional weights and the perceptual space (e.g. INDSCAL, Carroll and Chang, 1970). It may well be that Carlson and Granström's listeners have the same perceptual space underlying dissimilarity judgements for both sets of instructions, but are simply varying the weight they attach to each of the dimensions.

Other difficulties may not be so easily handled. Although the Dutch group's approach implies that vowel quality is essentially the same thing as timbre, there are effects of fundamental frequency that are difficult to explain in this way. In particular, while it is known that fundamental frequency can affect perceived vowel identity (in a way expected from the covariation of formant frequencies and fundamentals in men, women and children (e.g. Miller, 1953)), psychoacoustical timbre is claimed to vary little with changes in fundamental frequency. Plomp and Steeneken (1971), for example, present data for some nonspeech sounds showing convincingly that fundamental frequency and spectral shape (fixed in absolute frequency) do not interact in determining pitch and timbre. Slawson's (1968) studies on vowel quality and timbre judgements of the same vowel-like sounds can also be interpreted to support the notion that fundamental frequency changes cannot be traded off against spectral envelope changes for nonspeech as they can for speech (even though the changes in fundamental are much less important than changes in the spectral envelope for determining vowel quality). This is just one reflection of the fact that vowel identification and timbre must be differentiated at some level, due to the specific cognitive processing necessary for speech sounds. There need be no conflict between these sets of processes: the crucial thing is to determine the extent to which general auditory processes can account for the phenomena observed in speech, and where they are inadequate.

x Vowel quality and perceptual grouping

Much more fundamental criticism of a simple spectral shape (or formant extracting) model for vowel perception comes from the work of Darwin (1984). Noting that speech sounds in the real world are almost always heard in the presence of other sounds, he has tried to illuminate some of the mechanisms that enable listeners to determine which parts of the incoming acoustic stream belong to a single speaker, and which to some other source.

A typical experiment uses a continuum of short (56 ms) steady-state vowel sounds (with a constant fundamental of 125 Hz) in which F_1 alone is varied to span an /ɪ/ to /ɛ/ continuum. Then, various manipulations of the amplitude of the fourth harmonic are made by adding another component at the same amplitude, frequency and phase as the original harmonic. The effect of this manipulation on vowel quality is estimated by determining the phoneme boundary shifts between the original and manipulated continua.

(A phoneme boundary is simply the place on the continuum where responses are equally split between the two possible choices.)

When the added tone starts and stops simultaneously with the vowel, the increase in the amplitude of the fourth harmonic (which at 500 Hz is usually above F_1) makes the vowels more /ɛ/- like and hence shifts the phoneme boundary downwards. More interestingly, making the added tone start 240 ms before the vowel erases the effect—it is now as if the extra energy had not been added into the fourth harmonic. That this shift in vowel colour is not due solely to a frequency-selective adaptation (and hence an effective reduction in the amplitude of the physically increased harmonic) is demonstrated by making the added tone start simultaneously with the vowel, but now, continue 240 ms after the vowel finishes. This also causes the phoneme boundary to shift towards the value obtained with no manipulation, although not so completely.

According to Darwin, this occurs because making the extra tone lead or lag the rest of the vowel encourages its assignment to a different source of sound from the one emitting the vowel. Thus, the perceptual system can effectively 'subtract' its effect. When entirely simultaneous with the vowel, however, the evidence is strong that the increased harmonic 'belongs' to it. and hence that the vowel must be more /ɛ/-like. Clearly, no simple measurement of auditory spectra can account for these effects. Some perceptual grouping mechanism must precede the vowel quality determination. Nor are such processes specific to vowel colour judgements; Bregman and Pinker (1978) have shown similar results in the perception of nonspeech timbre. Interestingly, though, Darwin also presents evidence for grouping mechanisms that seem to be specific to speech.

xi The role of temporal information in spectral pattern perception

Although it has long been recognized that the time structure of a sound can affect timbre, even in the strict sense of the term we have used here (e.g. as in the work on phase relationships among the harmonic components of a periodic sound), this was considered of relatively minor import, and to result from a different mechanism than the one primarily responsible for timbre perception. In so far as any physiological connections were made (and psychoacousticians rightly tend to shy away from strong statements on this topic), timbre was seen to reflect 'the *distribution pattern* of the neural activity along the basilar membrane' (Plomp, 1976) while any structuring in time, arising from the interaction of components, was held responsible for the degree of 'roughness' of a sound.

As already discussed in Chapter 2, this view may not be tenable. Studies of firing patterns in the auditory nerve have indicated that the mean firing rates across fibres of varying characteristic frequency may not reflect the physical spectrum to the extent necessary to account for behavioural

performance, especially at high levels (Sachs and Young, 1979) or in background noise (Voigt *et al.*, 1981). As it is well known that there is a synchrony between auditory nerve impulses and the effective stimulating waveform (see Chapter 2), it has been suggested that temporal aspects of the firing patterns are used in the extraction of spectral information (Young and Sachs, 1979). Such analyses, which in effect combine rate, place and timing information, do seem to retain sufficient spectral detail (or possibly too much) in comparison to what can be inferred from human performance.

Models like these can, at least in principle, account for the effect of phase differences on timbre. As long as harmonics interact on the basilar membrane, phase changes can cause a change in the stimulating waveform, and hence a change in the temporal detail in the nerve firing patterns. This will presumably be reflected in the neural representation of spectral patterns computed by a mechanism which relies on temporal structure.

One interesting implication of these ideas is that variations in phase and spectral shape, although they may differ in the extent of their effect, are finally expressed in a unitary form. This is rather different from the views of the Dutch group, who have tended to see temporal effects (like those of phase) as arising from a mechanism distinct from that primarily responsible for timbre.

Unfortunately, there is relatively little psychoacoustical evidence bearing directly on the extent to which temporal factors are involved in timbre perception. We have already noted a number of experiments showing phase effects, but, at least in Plomp and Steeneken's (1969) work, phase changes seemed to result in variations along a different perceptual dimension from that related to spectral amplitude changes. Results more in line with the idea that phase and amplitude manipulations can influence perceptual judgements in similar ways come from Rosen (1984, 1986a,b), who studied phase manipulations in three-harmonic complexes of fundamental frequency 125 and 250 Hz. Some listeners (with both normal and impaired hearing) reported that the phase change caused a change that corresponded to a change in vowel colour, typically from /i/-like ('ee') to /u/-like ('oo'). The simplicity of such stimuli, as well as the important effects which result from what seem to be fairly minor changes in the stimuli (e.g. from using sine or cosine phase as the reference condition) may make them especially useful in physiological experiments. Using more complex stimuli, Darwin and Gardner (1986) have demonstrated that phase changes can alter the perceived first-formant frequency in vowels. In general, phase manipulations should be a powerful tool in elucidating temporally-based mechanisms of auditory analysis, as they leave the spectral amplitude pattern unchanged.

Possibly the strongest evidence for temporal coding of timbre comes from studies of deafened patients implanted with single electrodes. Such electrodes are placed either extracochlearly, on the round window or

promontory (e.g. Fourcin *et al.*, 1979) or intracochlearly through the round window into scala tympani (e.g. House, 1976). In either case, currents passed through the electrode stimulate the auditory nerve and lead to auditory sensations. It is fairly clear from theoretical considerations, psychophysical tests (Shannon, 1983) and physiological investigations (Kiang and Moxon, 1972) that there is no differential stimulation of nerve fibres according to their position in the cochlea—in other words, no place analysis. Hence, any timbre differences perceived by these patients must be due to purely temporal processes.

Given this fact, some extraordinary results have been reported in a small group of patients who have been stimulated with a frequency-weighted and amplitude-compressed version of the microphone-transduced speech signal (Hochmair-Desoyer *et al.*, 1985; Owens *et al.*, 1982). These so-called 'star' patients have exhibited a degree of understanding of continuous speech without lipreading that can only be explained if they are able to perceive some of the spectrum differences in speech. White (1983), examining the performance of one of these patients under well-controlled conditions, has shown that differences in first formant frequency are readily perceived, and can elicit the appropriate vowel labels. On the other hand, second formant changes are nearly imperceptible. Rosen and Ball (1986) have also found evidence for sensitivity to spectral changes in the F_1 region of vowels in two patients with the Hochmair device, even though they cannot understand speech by lipreading alone. It is almost certain that these timbre changes are perceived via mechanisms that operate purely on the temporal structure of nerve firing patterns, and more thorough investigation of the perceptual capabilities of these patients is likely to lead to insights about the role of temporal processes in normal hearing.

xii Static spectral pattern perception in the hearing impaired

If the width of auditory filters has any bearing on the representation of spectrum shape in the auditory system (which even the strongest proponents of temporal processes would not deny), then the reduced frequency resolution found in hearing-impaired listeners should greatly influence their ability to distinguish speech contrasts based on spectral differences. We will focus attention here solely on those studies involving the perception of vowel or vowel-related differences; these constitute the vast majority, in any case.

Consider what might be expected on the basis of a simple excitation-pattern model, like the one discussed earlier, when frequency selectivity is degraded. Figure 7.28 shows how the spectral envelopes of two natural vowels are represented at the outputs of two auditory filter banks: one with normal selectivity, and one with filter bandwidths increased by a factor of 3.

Clearly, much less detail of the spectral envelope is preserved when auditory filters have wider bandwidths; the peak-to-valley differences in the resulting

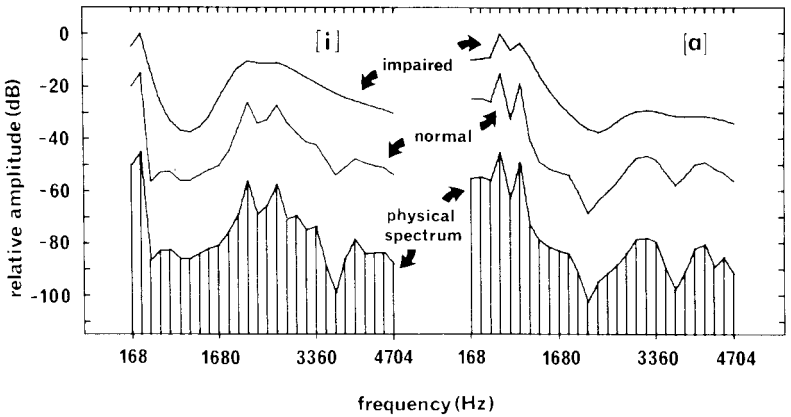


FIG. 7.28 The outputs of two hypothetical auditory filter banks to two synthesized vowels modelled closely on the continuously held utterances of an adult male. Both vowels have a fundamental frequency of 168 Hz, and the harmonics making up each vowel can be seen. Lines drawn between the peaks of each harmonic component define the spectral envelopes. To obtain the representations of the vowel spectra at the output of the filter bank, calculations were made of the total power passed through a set of 28 filters, the centre frequencies of which corresponded to the harmonic frequencies. For the 'normal' filter bank, the auditory filters were assumed to have an amplitude response given by a rounded exponential, and with the equivalent rectangular bandwidths given by Moore and Glasberg (1983b), as in Fig. 7.15. For the 'impaired' filter bank, filter bandwidths were increased by a factor of three, as in Fig. 7.16. The technique used here is similar to that used in the construction of excitation patterns (Chapter 5). The differences lie in the restriction of filter centre frequencies to those that coincide with a harmonic (for clarity in seeing the representation of the spectral envelope), and the use of a linear, rather than ERB or Bark, frequency scale.

excitation pattern are much smaller than normal. Note how the spectral valley between F_2 and F_3 in the [i] vowel (on the left of Fig. 7.28) has completely disappeared with the wide filtering. Overall, the effect is one of 'blurring' the spectrum. These effects should be directly reflected in excitation patterns derived from masking or pulsation-threshold experiments. They should also result in degraded performance in distinguishing changes of formant frequency (as this is assumed to depend on the slope of the excitation pattern). Two caveats are necessary before we go on to consider the data.

Firstly, in much of the relevant work, the frequency selectivity of the listeners has not been explicitly measured. As is discussed in Chapter 6, although most hearing-impaired listeners do indeed show reduced selectivity, not all do. Furthermore, there is a great deal of variability in selectivity for listeners with the same pure-tone thresholds, which is often the only quantitative information provided.

Secondly, there are other auditory abilities of the patients that may influence performance. For example, in the excitation-pattern model, a degradation in formant frequency jnd may be the result of impaired auditory filtering, or impaired discrimination of changes in excitation level, or

both. More confusingly, impaired auditory filtering could be compensated for by improved discrimination of changes in excitation level. Without further information, these two factors are hopelessly confounded.

With these warnings in mind, let us first examine the data relating to formant frequency resolution.

Pickett and Martony (1970) measured the ability of six listeners with severe binaural sensorineural hearing loss to discriminate changes in frequency of a single formant in the F_1 region (205–825 Hz) and compared their results to those of four normal listeners. The impaired listeners did surprisingly well, considering their very low scores in identifying monosyllabic words without lipreading (0–16%). Although they showed large learning effects, asymptotic performance was near normal for the low-frequency formants (205, 400 and 275 Hz) as long as the hearing loss in that region was moderate (<60 dB). At 825 Hz, where all the impaired listeners had at least an 85 dB loss, the best group had jnds of about 5–10% compared to the 3–5% of normal listeners. The group with the more severe loss (averaging about 100 dB) had jnds above 10%, which were attributed to tactile sensations. (Recall from Section 3Biii that we estimated formant frequency jnds of about 6% to be necessary for speech contrasts.)

That this discrimination task also reflected something about vowel perception in more realistic situations was shown by the outcome of another test. Here, listeners were required to identify natural consonant-vowel-consonant (CVC) monosyllables but were given a choice between two items in which only the vowel varied. (For example, if the test word were 'gauze', the foil might be 'goes'.) Six vowels were paired in all combinations and five word pairs were used for each contrast. Ranking the listeners according to their mean jnd for the higher three formants (from best to worst) gave the same order as the ranking resulting from performance on the forced-choice test using natural vowels.

Nakagawa *et al.* (1983) also found much larger jnds in frequency for formants at and above 800 Hz. They tested the ability of hearing-impaired and normal children to discriminate changes of second-formant frequency (spanning the range 800–2240 Hz) in a set of five synthetic five-formant vowels, modelled closely on natural speech. One of the four hearing-impaired listeners was unable to perform the task at all. The other three always had jnds larger than normal, the extent of the degradation in performance varying over a wide range with vowel and listener.

Results like these seem well accounted for, qualitatively speaking, by excitation-pattern models with widened auditory filters, at least when hearing-impaired listeners actually exhibit impaired performance. Horst and Ritsma (1982, also reported in more detail in Horst, 1982) have presented some data, however, that are difficult to account for completely with such a model. They measured psychophysical tuning curves (see Chapters 3 and 4) in a set of

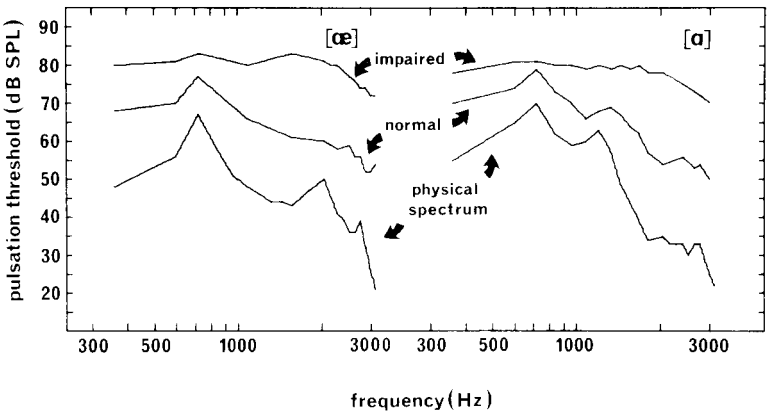


FIG. 7.29 Mean vowel pulsation patterns obtained from normal and hearing-impaired listeners for sounds at 100 dB SPL. The spectra of the vowels are set arbitrarily on the ordinate for clarity. The pulsations patterns from the normal listeners are more similar in shape to the physical spectra than those obtained from the impaired listeners. Redrawn with permission from Bacon (1979).

normal and moderately hearing-impaired listeners, and looked at the relationship between selectivity and frequency discrimination ability for the triangular-spectra, single-formant-like sounds discussed above. Two spectral slopes were used: 25 and 200 dB/octave. Recall that the excitation pattern model predicts better discrimination for the more steeply sloping 'formant' only when the auditory filters are sharper than 25 dB/octave, as is found in normal listeners. Impaired listeners, with greatly reduced selectivity, would be expected to show smaller, or no, differences between the two conditions.

All the impaired listeners did, in fact, show impaired selectivity. Furthermore, although there was wide variability, the impaired listeners showed a tendency, on average, for increased jnds with both spectral slopes used. As for the normal listeners, jnds tended to be smaller for the sounds with steeper spectral slopes. However, there was no significant correlation between the degree of frequency selectivity and the ratio between the frequency-jnds obtained for the shallow and steeply sloping sounds, which should decrease (becoming closer to unity) with worsening selectivity.

Much more striking differences between hearing-impaired and normal listeners have been obtained in attempts to measure the excitation pattern directly. Bacon and Brandt (1982) used the pulsation-threshold technique (Chapter 4) to estimate the excitation patterns of vowels for a set of normal and hearing-impaired listeners. As is seen in Fig. 7.29, the patterns derived from the normal listeners resemble more closely the physical spectrum of the vowels than those from the hearing-impaired, even at the same, relatively high, intensity level. Little or no spectral detail is preserved in the patterns

from the impaired listeners. In fact, we might suppose that they would not be able to distinguish the two vowels, a possibility not tested.

Striking as these results are, there are many problems in interpreting them. For one thing, one of the normal listeners exhibited highly abnormal vowel pulsation patterns. Although it is possible that this particular listener had impaired frequency selectivity, no measurements were made to check this. Correspondingly, it is not clear whether the abnormal patterns obtained from the impaired listeners were actually due to poor selectivity. Furthermore, there are considerable difficulties in interpreting pulsation threshold measurements in normal listeners (see Chapter 4); in impaired listeners, the difficulties have not even been considered. The extent to which independent measures of selectivity can be used to predict excitation patterns has yet to be determined. Finally, the impairment implied by such featureless vowel pulsation patterns seems much larger than would be expected from direct tests of vowel discrimination. Two of the impaired listeners, for example, had only moderate losses (<50 dB) over the frequency range of the first two formants in the vowels used. One would expect such listeners to be easily able to discriminate the two vowels, yet the pulsation patterns hardly differ.

Similar feelings have troubled Turner and Van Tasell (1984). Noting the essentially flat pulsation patterns measured by Bacon and Brandt, they concluded that peak-to-valley differences in the physical spectrum would have to be very large indeed for them to be represented at all in the pulsation patterns and hence in the internally represented spectra. Therefore, one should expect that hearing-impaired listeners attempting to detect the presence or absence of a spectral notch in a vowel-like sound would require much deeper notches for detection than normal listeners. To test this supposition, they synthesized a set of /æ/-type vowels with varying-depth notches at 2 kHz, and determined the just-detectable notch for three normal and two hearing-impaired listeners. One of the impaired listeners had a demonstrable loss of frequency selectivity in the frequency region of the notch, yet he could just detect a notch of slightly less than 2 dB, close to the average of the normal listeners. The other impaired listener (whose selectivity was not determined) required a notch just over 3 dB deep, worse than the average of the three normal listeners, but not markedly larger than the almost 3-dB notch required by the normal listener who performed worst. In any case, this notch is still small compared to those found in natural vowels.

Turner and Van Tasell then applied a simple excitation-pattern model to their stimuli and showed that even large variations in the bandwidth of the auditory filters led to surprisingly small changes in the predicted just-detectable notch. They concluded that auditory filtering would have to be nearly non-existent for the notches found in natural vowels to be undetectable, and hence that the vowel pulsation patterns of Bacon and Brandt probably do not accurately reflect the spectral shape information available to impaired

listeners. Of course, it may also be that Bacon and Brandt's listeners had much less selectivity than Turner and Van Tasell's; with such sparse information about both sets of listeners, it is impossible to know.

Clearly, much work remains to be done in this area. It is essential that any study which attempts to explore the perceptual consequences of reduced frequency selectivity must actually measure selectivity directly. Measures of intensity discrimination would also be desirable, since, as we noted above, variations in the degree of selectivity are often confounded with variations in sensitivity to intensity changes. Furthermore, more careful attention needs to be paid to the particular task used, and the extent to which changes in frequency selectivity will alter performance. As Turner and Van Tasell showed, the detection of a spectral notch is rather unimpaired by large variations in selectivity.

There is evidence that the accuracy of vowel identification, too, is relatively insensitive to variations in selectivity. Klein *et al.* (1970) used their simple filter-bank model of timbre perception to construct a vowel 'recognizer' and assessed its performance for utterances from 50 different young male adult speakers. The use of ten 1/3-octave bands (meant to approximate auditory filter bandwidths) between 500 Hz and 4 kHz gave 95% correct identifications, but cutting down the number of bands to three (two 1-octave and one 4/3-octave) in the same frequency range only reduced performance to 88% correct. This result may explain the general finding that hearing-impaired listeners make, as Owens *et al.* (1968) put it, 'a surprisingly low number of errors' in identifying vowels. In short, it is important to construct at least first-order models of the phenomena under investigation, in order to estimate the extent to which the frequency selectivity displayed by the listeners can actually account for the results obtained.

Finally, given current speculation about the role of temporal information in coding timbre in the auditory system (Section 3Bxi), it may well be that many of the results obtained from impaired listeners would be more easily explained with models that stress time, as well as place, information.

Confirming evidence for this point of view comes from some recent work investigating the perception of phase shifts by hearing-impaired listeners, already discussed in Section 3Av. There we noted that changes in relative phase can, under certain conditions, be better discriminated by the hearing impaired than by normal listeners. We also noted that some of these changes are described as subjective changes in pitch, but they are also often described as changes in timbre, specifically as changes in vowel colour (Rosen, 1984, 1986b). For example, one patient, with a severe hearing loss (varying from 60–100 dB HL over the range 0.25–1.0 kHz), was presented with three-harmonic complexes of fundamental frequency 250 Hz. Shifting the phase of the 500-Hz component by 90° caused a change in percept from 'ee' to 'oo'. Since hearing-impaired listeners tend to show a decreased sensitivity

to changes in spectrum shape, and an increased sensitivity to phase changes, it seems likely that the relative importance of phase is much greater for them than for normal listeners. Studies of hearing-impaired listeners along the lines of those performed by Plomp and Steeneken (1969) using normal listeners may reach very different conclusions about the role of phase.

This sensitivity to phase changes may also be a factor in the difficulties many hearing-impaired listeners have in reverberant environments, apart from affecting the saliency of voice pitch (Section 3Av). Although the randomization of phase relationships caused by reverberation has little effect on the timbre perceived by normal listeners (Plomp and Steeneken, 1973), hearing-impaired listeners might well be affected more severely.

C *Dynamic patterns of spectral change*

Since the shape of the vocal tract, for a given excitation source, is the primary determinant of the output spectrum, the spectrum will change whenever the vocal tract is moving (and even when it is not, if the excitation changes). As most speech is characterized by movements of the vocal tract (!), many would consider patterns of spectral change to be more important than static patterns in accounting for normal speech perception. But, as often happens, we have the least to say about the most important things. Although there is a fairly good understanding of the physical nature of the dynamic patterns found in speech, and the phonetic contrasts they contribute to, there is relatively little known about the psychophysics of the perception of such patterns, and even less about the influence of frequency selectivity. Of necessity, then, our exposition will be considerably shorter than that in the previous sections.

i Importance in speech contrasts

Befitting their central role in speech, dynamic patterns of spectral change contribute to contrasts in all of the three major features traditionally used by phoneticians: voicing, manner and place.

Although there are many voicing distinctions which depend on relatively steady-state properties (e.g. for fricatives), at least two cues to some plosive voicing contrasts are essentially dynamic (Section 2B). In general, voiced plosives are distinguished from voiceless ones by the timing of the onset of vocal fold vibration relative to the oral release. In languages like English and German, voiced plosives have vocal fold vibration starting more or less simultaneously with the release; in voiceless plosives, there is a delay, filled by aspiration noise, between the release and voicing onset. When this aspiration noise excites the vocal tract, the first formant will be poorly represented in the auditory system for two reasons. Firstly, the excitation is fairly weak. Secondly, the relatively flat spectrum of the aspiration tends to

excite all formants equally, but auditory sensitivity is greater in the F_2 and F_3 frequency region than it is in the F_1 region. Voiceless plosives initially exhibit spectra with an effective high-frequency emphasis which changes to a low-frequency emphasis at the onset of voicing, because the laryngeal source has a spectrum weighted towards low frequencies. Voiced plosives have this low-frequency emphasis at or close to release continuing into the following vowel. This gross spectral change at lower frequencies associated with the appearance of F_1 at high amplitude has long been known as F_1-cutback, since, as the voice-onset time increases, F_1 is 'cut back' more and more (Liberman *et al.*, 1958).

There is another dynamic spectral feature associated with the interaction between voicing onset and F_1. If the vocal tract comes to a relatively steady-state position before the onset of voicing—when the first formant makes a strong appearance—then the first formant, too, will be relatively steady state. If, however, the voicing appears soon after the plosive release and while the vocal tract is still moving, the voice-excited first formant will be rising. Hence, the presence or absence of a first-formant transition (in addition to its time of appearance) influences the perceptual decision about whether a plosive is voiced or voiceless. Soli (1983) has claimed that a cue related to this feature (the F_1 onset frequency) is primary in determining the perceptual responses of infants, adults and animal listeners.

Manner contrasts are even more strongly dependent on dynamic patterns of spectral change. Probably the major cue resides in the rate at which the spectrum is changing. At the most basic level, Stevens and Blumstein (1981) claim that rapid spectrum changes are an important feature distinguishing consonantal from nonconsonantal sounds (like vowels and glides). Similarly, diphthongs have fairly slow transitions, while semivowels like /j/ (*y*elp) and /w/ change rather more quickly. Plosives exhibit the fastest changes. So, for example, it is possible to create a /bɛ/ to /wɛ/ continuum simply by varying the rate of the formant transitions which signal the movement of the vocal tract from a bilabial closure to a position appropriate for the vowel (Liberman *et al.*, 1956). However, Shinn and Blumstein (1984) have recently argued that the amplitude envelope at release is a much more potent cue to this contrast.

Finally, the dynamics of spectral change are also intimately connected with place of articulation. Here, the primary feature seems not to be in the rate of change, but in the details of the spectral movements. For example, one important cue to the distinction between syllable-initial plosives like /bɑ/ and /dɑ/ or /pɑ/ and /tɑ/ is the direction of the second formant transition (whether excited periodically or aperiodically), which is generally rising for the bilabials and falling for the alveolars. In general, formant transitions can be important cues for distinguishing place of articulation in plosives (Delattre *et al.*, 1955) and nasals (Liberman *et al.*, 1954; Nord, 1976).

Even some place distinctions among the fricatives rely heavily on formant movements, as opposed to steady-state patterns, for reliable identification. This was shown by Harris (1958), who recorded a set of fricative-vowel sequences (e.g. /fi/, /θi/, /si/, /ʃi/) and cross-spliced the frication noise of each with the vocalic portions of all the others. Although spliced syllables beginning with relatively high-intensity /s/ and /ʃ/ fricative noises were largely identified as /s/ and /ʃ/ respectively, no matter what the following vowel, syllables beginning with /f/ and /θ/ frication tended to be identified on the basis of the vocalic portion, with its formant transitions.

The movements of formants are also important in distinguishing place of articulation among the dipthongs, and glides (in English known as the 'whirly' set /w/, /r/, /l/ and /j/; O'Connor *et al.*, 1957). Furthermore, some workers have proposed that dynamic formant patterns are crucial even in the perception of vowels, a point of view advocated by Shankweiler *et al.* (1977), although there is much debate about this issue (Howell, 1981; Pisoni *et al.*, 1979).

ii Major acoustic features

Over short enough stretches of time, we can think of dynamic patterns of spectral change as a series of static spectra. Hence, within these short time intervals, the relevant acoustic features will be more or less the same as those we have already described in Section 3Bii (in terms of formant frequency ranges and bandwidths). The only further information needed is an indication of the rates of change that are found.

In fact, Liberman *et al.* (1956) claim that the important perceptual feature is not the rate of change, but simply the duration over which the transition extends. They conclude this from the results of a study in which listeners were asked to categorize simple synthetic syllables as either 'b-vowel' or 'w-vowel' for six different vowels and a range of transition durations. As the initial starting frequencies for the two formants were always the same, and the steady-state formants differed appropriately for each of the vowels, the rate of frequency change differed for different vowels. The resulting categorization curves overlapped considerably when plotted as a function of transition duration, but were quite disparate when plotted as rate of formant frequency change.

In a similar study using a single vowel with initial transitions that varied over a wider range of duration (10–300 ms), Liberman *et al.* (1956) found that a transition–duration continuum could be divided by the listeners reasonably well into three classes. The shortest transition durations ($<$60–90 ms) were heard as plosives (/bɛ/ or /gɛ/), intermediate durations (about 100–200 ms) as glides (/wɛ/ or /jɛ/), and the longest ($>$200 ms) as 'vowels of changing color' (/uɛ/ 'oo-eh' or /iɛ/ 'ee-eh').

These values may not, of course, reflect accurately the transition durations found in natural speech, and they are certain to shift with changes in factors such as speech rate (Miller and Baer, 1983). Detailed measurements of spectral changes are rare, not least because of the difficulties in precisely determining, in both frequency and time, where transitions begin and end (e.g. see Halle *et al.*, 1957). Some general comments can be made, however.

Diphthongs, the dynamic sounds that exhibit the slowest changes over time, often show continuous changes in spectrum over their entire duration. In the studies of Holbrook and Fairbanks (1962), using words like 'hay' and 'high' spoken in a constant carrier phrase, this was of the order of 200–250 ms. Even with such slowly changing sounds, formant frequencies could change at average rates of nearly 5 octaves/second (F_2 in 'Oi').

Glides and plosives typically show faster changes. Pickett (1980), for example, claims that the formant transitions associated with glides last about 75–100 ms while those for plosives last about 50 ms. Mack and Blumstein (1983) measured a set of /w/- and /b/-vowel utterances from two speakers and found F_1 transitions to last an average of 41.5 and 23.4 ms, and F_2 transitions, 75.3 and 26.1 ms, for /w/ and /b/ respectively. This can lead to formant frequencies changing as fast as 18 octaves/second (F_2 in /wi/).

Sometimes, the speech spectrum can change faster still. The vocal folds can go from inactivity into full vibration within two to three vibratory cycles. At the junction between nasals and vowels, the nasal cavities can be coupled or uncoupled by very quick movements of the velum, so that the appearance or disappearance of dips in the spectrum due to antiresonances, or the increase or decrease in formant damping, can take place in the interval from one voice pitch period to the next. Plosive release bursts can cause the spectrum to change drastically over the course of a few milliseconds.

iii The effect of time-bandwidth limitations on the representation of dynamic spectral patterns

As we noted earlier, not all of the acoustic detail present in speech is perceivable by the human listener. At least one constraint on the joint perception of changes in frequency and time may be attributed to relatively peripheral aspects of auditory processing. We have seen evidence in Chapter 5 that the narrow bandwidths of low-frequency auditory filters (below about 1–3 kHz) put limits on the accuracy with which temporal detail can be registered. Hence, phonetic contrasts which rely primarily on temporal detail will be best perceived through high-frequency auditory filters (for example, detecting a silent gap which can cue the presence of a voiceless plosive as in 'speed' *vs* 'seed'; Bailey and Summerfield, 1980).

On the whole, there has been little explicit consideration of such constraints in the speech perception literature, even by those theorists who posit the primacy of acoustic/auditory factors. For instance, Stevens and Blumstein

(1981) have explored the idea that an invariant cue for place of articulation in plosives can be found in the spectrum shape within the 10–20 ms following consonantal release. They have argued that the initial stages of the formant transitions (whether periodically excited as in voiced plosives, or aperiodically excited as in voiceless ones) are integrated with the release burst, to form an essentially steady-state cue. The formant transitions themselves are supposed to be secondary, serving more to link the acoustic cues at plosive release to those of the following vowel, than to provide primary cues to place of articulation.

This notion, that spectral energy is smoothed over time, seems eminently sensible given the finite bandwidths of auditory filters. It may well be that using integration time windows that vary appropriately with frequency (rather than using a constant time of 25.6 ms as Stevens and Blumstein did) would more accurately reflect the features of the spectrum available to a human listener.

Similar ideas may be applicable in accounting for the auditory representation of the cue to plosive voicing found in the first formant. Stevens and Klatt (1974) claim that the important factor is the presence or absence of a rapid spectrum change (a dynamic feature), while Soli (1983) has focussed attention on the spectrum at voicing onset (a static feature). At least part of the resolution of this controversy is likely to be found in a consideration of auditory filtering. For a formant of a given bandwidth and transition rate, the lower its frequency, the more it will be represented in the auditory system as a single quasi-steady-state spectrum, due to the time-smearing effects of the relatively narrowband low-frequency auditory filters. Conversely, the higher in frequency it is, the more likely it will be represented as a formant changing its frequency in time.

iv Models for the perception of dynamic spectral patterns

Considerations like those discussed in the previous section have played very little part in theorizing about the way in which listeners use dynamically varying spectral information. Discussions have centred more around the nature of the process that samples the dynamic spectrum; for example, whether or not the sampling is uniform in time, and if not, how it is controlled (Bladon, 1984). Yet the properties of auditory filters must play an important role in determining the sensitivity of human listeners to these kinds of acoustic features. Perhaps the most hopeful sign is that many researchers are constructing models of the peripheral auditory system, based on either psychoacoustic (e.g. Chistovich *et al.*, 1974), or physiological (e.g. Delgutte, 1982), findings. These are being used to process speech signals, either for studying the acoustic correlates of phonetic and phonemic categories, or as first stages for speech recognizers. In so far as auditory filtering is accurately modelled, the limitations on temporal resolution ascribable to peripheral filtering will also be included.

v Performance of normal listeners

Although there are few studies, some general points can be made. Firstly, formant frequency jnds are rather larger for dynamic as opposed to steady-state sounds. This was first shown for speech-like sounds by Mermelstein (1978) who used five-formant syntheses modelled on either consonant-bound or steady-state vowels. Interestingly, even when F_2 was held stationary and only F_1 and F_3 varied in time, the jnd for F_2 was larger than that obtained when all formants were steady-state.

Horst (1982) measured jnds in centre frequency for static and dynamic sounds resembling single formants, and also obtained superior discrimination with steady-state stimuli. Furthermore, he has shown that the just-detectable formant-frequency sweep varies inversely with its duration, and that, just as for stationary signals (Section 3Biii), jnds decrease with decreasing formant bandwidth. However, this variable has less influence for dynamic than for steady-state sounds. Horst attributes this to the spectral spreading that occurs with dynamic sounds, giving them all effectively wider spectra. This spread may also partially explain the general finding of worsened performance with dynamic patterns, since all such sounds can be considered to have their bandwidths widened by being swept in frequency.

vi The perception of dynamic spectral patterns by the hearing impaired

Ironically, although the amount of work on the perception of dynamic spectral patterns by normal listeners is small, quite a considerable body of data from hearing-impaired listeners is available. This is probably due to the fact that such acoustic features are crucial in speech perception, yet they are often unusable by impaired listeners.

Nearly all of the work that may be related to auditory frequency selectivity has involved the perception of second formant transitions. The early studies were performed by Pickett and his colleagues and receive an excellent review in Pickett *et al.* (1983). Typically, the minimum frequency change necessary for a transition in F_2 to be detectable (compared to a formant with no transition) is determined when F_2 is presented on its own, and in the presence of a first formant. At comfortable levels, the presence of F_1 has little effect on the just-detectable transition for normal listeners. For many impaired listeners, however, the presence of the first formant causes the just-detectable second formant transition to increase, often markedly. Danaher *et al.* (1973), for example, found that in a group of listeners with sloping audiograms, the average just-detectable transition increased by a factor of two from what was already a very impaired performance with F_2 alone.

This degradation of performance was attributed to excitation from F_1 spreading upwards and masking the information conveyed by F_2. Further experiments buttressed this point of view. Danaher *et al.* (1973) found that

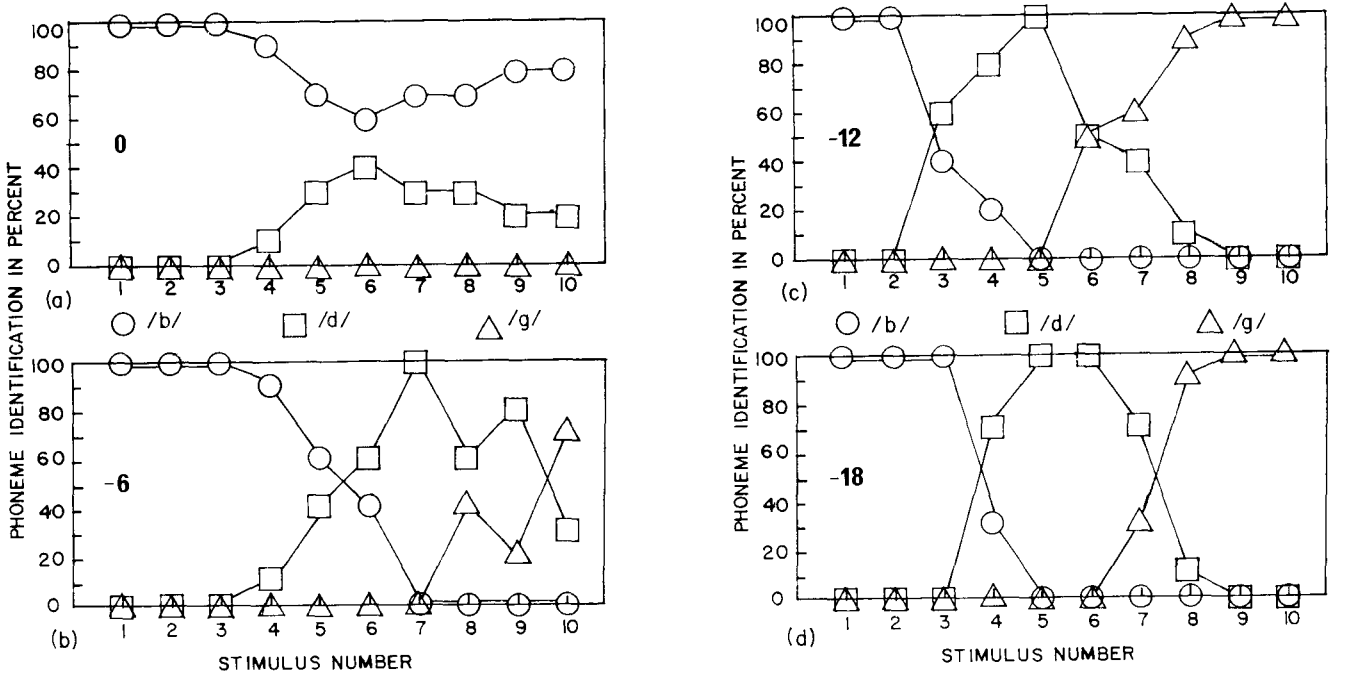


FIG. 7.30 Categorization functions obtained from a single listener with a demonstrated loss of frequency selectivity. The stimuli formed a two-formant, voiced-plosive, place-of-articulation continuum /bɑ/–/dɑ/–/gɑ/. The first formant was either at the same level as the second formant [panel (a)], or 6, 12 or 18 dB less intense [panels (b), (c) and (d), respectively]. Note how the reduction in first-formant amplitude leads to the emergence of normal categorization functions. Reproduced with permission and minor modifications from Hannley and Dorman (1983).

the just-detectable transitions measured in *normal* listeners could be increased in the presence of F_1 if the stimuli were presented at amplitudes more similar to those used for the hearing impaired. This is easily understandable given the decrease of the low-frequency slopes of auditory filters with increases in level (Chapter 3). It was also shown that the performance of impaired listeners for two-formant stimuli could be improved if the amplitude of F_1 was reduced (Danaher *et al.*, 1973) or if it was presented to the other ear (Danaher and Pickett, 1975).

Danaher *et al.* (1973) and Danaher and Pickett (1975) suggest that the degradation of performance exhibited by the hearing-impaired listeners in the presence of F_1 occurs not because their frequency selectivity is impaired, but because they need to listen at levels so high that even the selectivity of *normal* listeners is impaired. Hannley and Dorman's (1983) findings are more in agreement with the view that reduced frequency selectivity, *per se*, can be responsible for an F_1-masking effect. Here, the existence of degraded frequency selectivity was confirmed by a masking paradigm in a set of impaired listeners with more moderate losses than those used in the studies discussed above. The same listeners were tested in an identification task using a set of /bɑ/–/dɑ/–/gɑ/ continua varying in the onset frequency of the second formant (and hence in the direction and extent of its transition to the unvarying steady-state vowel). Across continua, the amplitude of the first formant (which always had an initial rising transition) was varied from equality with the second formant, to a condition in which it was 18 dB lower. At least some of the impaired listeners showed a large improvement in performance as F_1 was decreased in amplitude. Figure 7.30 shows the results from one listener with a very minor hearing loss in the F_2 region (about 10 dB) who nevertheless showed a marked susceptibility to the upward spread of masking in the separate masking study.

F_1 attenuation did not improve performance for all listeners. On the whole, presbyacusic listeners exhibited the least effect while listeners with noise-induced hearing loss were helped the most, even though the noise-masking task did not show any differences in frequency selectivity. Clearly, other factors must be at work. It is possible, for example, that the two groups differed widely in their ability to hear F_2 transitions in isolation. The presbyacusics might not have been able to detect the presence of a transition even if F_2 were presented on its own, and so it would not be surprising that their performance showed little improvement as F_1 was attenuated. Indeed, their performance was almost uniformly worse than that of the other groups, even when F_1 was attenuated by 18 dB. There are almost certainly a number of other auditory capacities called into play in this task, as Hannley and Dorman note (e.g. temporal resolution) and a complete explanation of listeners' performance must take them all into account.

In short, some impaired listeners can benefit from attenuating low-frequency formants, and this (in addition to the commonly found increasing loss with increasing frequency) is probably responsible for the typically rising amplitude-frequency response recommended for hearing aids. Whether or not such a characteristic really is as universally appropriate as has been stated is a question we will address later (Section 4E).

Dorman and his colleagues now argue that, in more natural speech, the masking effect of the first formant is *not* an important factor in the difficulties faced by impaired listeners. For example, Dorman, Dobbins, Hannley and Lindholm (unpublished manuscript) synthesized versions of the voiced-plosive/vowel combinations /bɑ/, /dɑ/ and /gɑ/ in which both F_2 and F_3 transitions could be used to distinguish among the three sounds. For neither /dɑ/ nor /gɑ/ did a 6-dB reduction of the first formant amplitude aid identification of the stimuli (and correct performance with /g/ seemed more dependent on having an appropriate release burst). Yet for /bɑ/, about a third of the listeners (all those who had difficulty in identifying this sound), were significantly helped by a 6-dB attenuation of F_1. Since, of the three sounds, the second formant transition for /bɑ/ starts at the lowest frequency, we should expect its identification to be most impaired through masking by F_1. Unfortunately, Dorman *et al.* did not measure frequency selectivity in their listeners, and simply on the basis of the audiograms (known not to be reliable indicators of frequency selectivity—see Chapter 6) concluded that it was not possible to determine which listeners would be affected by the spread of masking.

To counter the argument that even /dɑ/ and /gɑ/ would have been identified more accurately with more extreme reductions in the amplitude of F_1, Dorman *et al.* (1985), in addition to full-formant sounds (versions of 'bay', 'day' and 'gay'), also synthesized stimuli without any first formants at all. Release bursts were also included in the synthesis, and it was found that at levels near those that led to maximum performance, eliminating F_1 did not improve identification accuracy. Only at levels much higher, and perceived as very loud by the listeners, did this occur.

Dorman *et al.* concluded that the (imputed) frequency selectivity of impaired listeners like the ones in this study was sufficient to overcome the upward spread of masking from F_1. What they neglect to emphasize is that the second formant in the 'bay'–'day'–'gay' sounds is considerably farther from the first formant than in the /bɑ/–/dɑ/–/gɑ/ sounds (about 1.8 octaves in the former and 0.7 octaves in the latter). It would be surprising, indeed, if F_1 did not affect the perception of /bɑ/ more than it did 'bay'. Furthermore, masking does not happen only from low frequencies to higher ones, and the auditory filters of hearing-impaired listeners do not necessarily only widen downwards (Chapter 6). Hence, F_3 may have been doing some masking in the plosive-'ay' stimuli as it was only 0.55 octaves from F_2,

whereas in the previous studies it was either absent, or an octave away.

In short, these results are well accounted for by assuming that hearing-impaired listeners (even of 'mild to moderate' degree) do have some degradation in frequency selectivity (but not a total one), which, under the right circumstances, results in some parts of the speech signal masking others. At times, it may be advantageous to decrease the amplitude of these masking components, or minimize their effects in other ways. The extent to which such manipulations are beneficial in real-life situations will be addressed below (Sections 4A, C, E).

Clearly, much work remains to be done in this area. A consideration of the possible masking effects from upper- as well as lower-frequency energy is likely to prove important, especially as hearing-impaired listeners can show asymmetries in their auditory filters (Chapter 6). Also, as we have noted before, information about the actual frequency selectivity of the listeners is crucial. Apart from these intraspeech masking effects, it seems likely that many of the problems encountered by the hearing impaired in perceiving dynamic spectral patterns are less related to deficits in frequency resolution than to deficits in temporal resolution. To go beyond this we need a better understanding of the relation between frequency selectivity and the perception of dynamic spectral changes by both normal and impaired listeners.

D *Masking of speech by extraneous sounds*

In all our discussions so far we have implicitly assumed that the signals reaching someone's ear are only those sounds the listener (or experimenter!) desires. Yet in everyday life, most speech is heard in a more or less intense background of other sounds, and it seems likely that frequency selectivity is as important in reducing the influence of these unwanted sounds as it is in processing speech itself.

The presence of a frequency selective mechanism, which essentially limits the interaction of different portions of the spectrum, has several advantages. For interfering noises with energy that is unevenly distributed across the spectrum, masking effects will tend to be isolated to the particular frequency region where the noise is most intense. Even in wideband noises with fairly flat spectra, spectral variation in the signal means that frequency regions where the signal is relatively intense compared with the noise can be isolated from regions where the signal is swamped by the noise. For a given signal energy, periodic sounds should be most resistant to masking, as their energy is concentrated at specific frequencies; hence auditory filters can include all of the energy in a single harmonic component while excluding much of the noise. In a situation like this, the narrower the filter, the better. When filters are wide enough to include several harmonic components, it might be possible

for the periodic fluctuations of beating harmonics to provide a temporal cue to aid detection. These are, of course, general principles, but what might we expect for human speech perception, given what is known about auditory filtering and the acoustic structure of speech?

First of all, if the degree of auditory filtering is crucial in determining the masking effects of noise on speech, then for a given ratio of speech-to-noise power, performance should worsen at higher levels since auditory filtering is broader there (Chapter 3).

More importantly, phonetic contrasts that are acoustically expressed in the low frequencies should be relatively resistant to noise, for three reasons. Firstly, the low-frequency components of speech tend to be more intense than those at high frequencies. Studies of the long-term spectrum of speech show it to be relatively flat between 100 or 200 and 600 Hz, then decreasing with a slope of about 10 dB/octave out to 10 kHz (Licklider and Miller, 1951). This shape is basically a reflection of the shape of the glottal source spectrum (see Section 2B) which is the primary energy source for speech. Secondly, at moderate to high sound levels, auditory filters are typically asymmetrical, with the slopes of their skirts being rather shallower on the low-frequency side. Thus, noise in low-frequency regions is much more effective at masking information at high frequencies than the other way around. Thirdly, auditory filters are narrowest at low frequencies, and so, at least for voiced sounds, can effectively isolate single harmonics from the surrounding noise.

For these reasons, we should expect the perception of vowel contrasts (especially those based on formant differences at low frequencies) to be generally more resistant to noise than the perception of consonantal contrasts. Within the class of consonants, however, there should be a wide variation in sensitivity to disruption by noise, depending upon the acoustic features of the particular consonant. Contrasts dependent on the presence or absence of voicing should be best preserved, as the phonetic feature of voicing is cued by a quasi-periodic low-frequency-rich signal. Furthermore, the quasi-harmonic spectrum due to larynx activity typically extends up to 3–4 kHz, so its acoustic reflection is spread over a wide band of frequencies (even though more intense at the low end) and may be perceived through any spectral 'holes' in the masking noise. Nasality should also be well preserved, at least in languages like English where nasals are typically voiced, as they too tend to be signalled by high-intensity low-frequency energy.

Place distinctions, when they are dependent, as they often are, on aperiodic energy which is spectrally diffuse, will be relatively greatly affected. In plosives, for example, this will cause one major cue, the aperiodic release burst, to be relatively susceptible to noise masking, as it is also of brief duration. Identification of place by the release burst should, however, be more noise resistant in voiceless plosives than in voiced ones, since voiceless plosive release bursts tend to be of greater intensity than those associated

with voiced plosives. Also, when in intervocalic position, voiceless releases occur after a longer silent interval, and this may also enhance their detectability. On the other hand, in so far as formant transitions are used to contrast different places of articulation, the voiced plosives, with their quasi-periodically excited transitions, will be better identified than voiceless plosives. Finally, among the suprasegmental features, voice pitch, like voicing itself, is likely to be robust, as will be the aspects of rhythm and syllabicity that are related to voicing.

There is also a temporal aspect, independent of frequency selectivity, to be considered. Contrasts made in relatively long-lasting and stable acoustic segments are likely to be more resistant to masking noise than contrasts based on quickly changing short segments, both because human listeners can integrate information over time, and because masking noises vary from moment to moment in frequency content and amplitude. Advantage may then be taken of dips in noise energy (Buus, 1985; Fastl, 1976, 1977a,b; Moore, 1985; Zwicker and Schorn, 1982). Although this temporal factor puts vowels at an advantage compared to consonants, some consonantal contrasts can be partly carried by vowel duration differences (see below). Part of the sensitivity to noise masking of place information (cued either by formant transitions or release bursts) may be accounted for by the brevity of its acoustic features.

i Performance of normal listeners

In general, studies on the masking of speech by noise bear out the expectations described above. One significant exception comes from the earliest work, which investigated the intelligibility of speech as a function of the speech-to-noise ratio for a white masking noise. Licklider and Miller (1951) claim that this remains constant over a wide range of absolute levels, a surprising result given the widening of auditory filters with increasing level. The main study cited there was by Hawkins and Stevens (1950) who measured the 'threshold of intelligibility' for speech, which was defined as 'the level at which the listener is just able to obtain without perceptible effort the meaning of almost every sentence and phrase of the connected discourse'. Once the sensation level of the wideband masking noise reached about 40 dB, 10-dB increases in noise level necessitated about 10-dB increases in signal level, out to the highest levels of the noise used (90 dB SL). There was no indication that the speech-to-noise ratio at the threshold of intelligibility increased with level. Whatever the reasons for this (and the vagueness of the criterion may play a role), Hawkins and Stevens found essentially the same results for the detection of tones in noise, whereas most studies have reported the necessary signal-to-noise ratio to increase slightly with overall level, as would be expected from the widening of auditory filters (see Chapter 3).

Later studies have taken a much more analytical approach, seeking to identify those features of the speech signal which are particularly susceptible to masking, and those which are relatively resistant. Almost all have tested the perception of consonants, as it is argued that they are both most important for intelligibility, and much more affected by noise than the vowels. The first, and still most influential of these, was performed by Miller and Nicely (1955). They presented a set of natural utterances of the form /Cɑ/, where C was one of 16 consonants, in the presence of various levels of wideband random noise. Analyses based primarily on articulatory features showed voicing to be most resistant to noise masking, followed closely by nasality. Place contrasts were most drastically affected by noise.

Even though such broadly defined articulatory features (in the sense of the range of acoustic correlates they encompass) give a reasonable account of the data, even more explanatory power is gained by dealing with more narrowly defined acoustic features. For example, although the perception of place of articulation is most susceptible to noise masking, we suggested above that this might be less so for voiced than voiceless plosives because the formant transitions are voiced in the former. Such is indeed the case, as a sub-experiment of Miller and Nicely shows (Table II). Here only the six English plosives were used as the consonants. Few voicing errors were made, only 6% overall, while place errors were rife at 54% of the responses. This much an analysis based on a system of independent articulatory features can account for. But if we look within the classes, we see that the place feature for the voiced plosives is correctly identified 51% of the time, whereas that for the voiceless plosives is identified only 37% of the time. Clearly, a more detailed analysis of the acoustic structure of the speech sounds involved would lead to an even better understanding of the structure of the confusion matrix. For example, among the voiced plosives, it appears that the bilabial /b/ is more easily distinguished from /d/ and /g/ than they are from each other. Presumably, this results from the fact that the acoustic cues for place in /b/ tend to be lower in frequency than they are for /d/ and /g/.

Interestingly, Miller and Nicely discussed this difference in the accuracy with which place of articulation was identified in voiced and voiceless plosives, and attributed it to the acoustic differences between the two classes of sounds. They also noted their particular dissatisfaction with the linguistic place feature, in essence because the acoustic features which were correlated with it were so disparate. In fact, although they discussed such matters under the heading 'Linguistic Features', it is clear that they were also trying to make their features reflect the acoustical structure of speech. (Note their feature 'duration', which is meant to distinguish the set /sʃzʒ/ with high-intensity long-duration aperiodicity from /ptkfθbdgvðmn/.) Much recent work on the confusions made between consonants (in particular by the hearing-impaired)

TABLE II
Confusion matrix based on Table XVIII of Miller and Nicely (1955)

Stimulus	Response p	t	k	b	d	g
p	37	18	36	4	3	1
t	25	34	35	3	1	2
k	27	28	40	1	2	1
b	4	3	3	68	14	8
d	1	1	1	13	53	31
g	1	4	3	15	48	31

These results were obtained in an experiment where relatively wideband speech signals (200–6500 Hz) were presented to listeners in a background of noise that was 12 dB more intense than the speech. The stimuli were presented as consonant-/ɑ/ syllables. The raw data of Table XVIII were converted into percentages (by row) to correct for the fact that the number of times each stimulus was presented varied slightly. Hence, each cell represents the percentage of times that the row stimulus was identified as the column stimulus, rounded to the nearest unit.

has not been so careful, and linguistic features are often used without caveats. Much explanatory power is lost this way.

One point to keep in mind when dealing with speech perception in the presence of noise is the likelihood that the relative importance of various acoustic cues will change from that found in quiet, and will vary with different types of noise. As we have been at pains to point out all along, any particular phonemic contrast can be expressed by a number of acoustic differences, and so it seems likely that those cues that are relatively resistant to masking will take on more importance in noisy environments. A good example of this is provided by Wardrip-Fruin (1982, 1985) who investigated the relative importance of cues to voicing in final plosives (e.g. 'bead' *vs* 'beat' or 'dig' *vs* 'dick'). It is well known (at least for words in isolation or prepausal positions) that the duration of the preceding vowel is considerably longer when followed by a voiced plosive than by a voiceless plosive (Umeda, 1975). There has been a great deal of controversy surrounding the perceptual importance of this cue in comparison to the cues found in the formant transitions related to the closure phase of the final plosives. What Wardrip-Fruin has shown is that the duration cue, of relatively little importance in quiet, can become the main cue in a background of noise. For example, when words with final voiced plosives had most of the vowel deleted, leaving the final transitions intact, the final plosives were still identified as voiced 82% of the time in quiet. Presented in a noise background, however, the same sounds were now judged to have a voiceless final plosive 87% of the time.

Relatively little experimentation has been done on the extent to which noise interferes with the perception of prosodic features. It has long been recognized (Miller and Nicely, 1955) that, due to the overall spectral shape of speech

(and also because of the way auditory filters vary in bandwidth across frequency), masking with flat-spectrum noise has perceptual effects very similar to those caused by low-pass filtering. Since the fundamental frequency contour is well preserved in low-pass filtered speech (as long as the cut-off frequency is above the frequency of the fundamental itself), it has always been felt that fundamental frequency is relatively resistant to masking.

Scheffers (1984) supports this assumption. He claims that changes in fundamental frequency of about 5% (probably sufficient for most linguistic contrasts, see Section 3Aiii) are just-perceptible when vowel-like sounds are just above masked threshold in pink noise. Unfortunately, his stimuli were rather unnatural from the point of view of speech, constructed in such a way that the spectral envelope shifted with changes in fundamental frequency. Listeners could have used this cue for discrimination, instead of the change in fundamental. Evidence that changes in the fundamental alone could lead to such good performance is found in Scheffers's results with pulse trains, which had relatively flat spectra, and in which this extra cue could not be effective. Here, changes of 5% in fundamental frequency could be discriminated as long as the sounds were about 5 dB above masked threshold.

Hoekstra (1979) also presents some data relevant to this problem. He determined jnds in fundamental frequency for periodic pulse-trains passed through a 1/3-octave filter at 2 kHz (as in the studies discussed in Section 3Av) and mixed with noise passed through the same filter. (Although these stimuli may seem remote from speech, the task given the listener is similar to the problem of perceiving the fundamental frequency of an utterance on the basis of one or two formants in the high-F_2/low-F_3 region when masking noise swamps the signal in other frequency regions.) The results, expressed as curves relating the jnd in fundamental frequency to the signal-to-noise ratio (referred to the level of noise necessary to just mask the signal at 40 dB SL), fell into two main groups depending upon the fundamental.

For high fundamental frequencies (about 150–200 Hz, depending upon the listener), where the harmonics are presumed by Hoekstra to be resolvable, the dependence of performance on noise level was similar to that found for a pure tone. The increase in the jnd with increasing noise level was fairly uniform (about a factor of 2 for each 3 dB) and changes of a semi-tone (6%) were discriminable even when the signal-to-noise (S/N) ratio was very poor (near unity). For fundamental frequencies lower than this, however, where several harmonics would be contained in the same 2-kHz auditory filter, the noise tended to have a much more abrupt effect. As the S/N ratio decreased to about 10 dB, performance worsened smoothly from what it was in quiet, at about the same rate as it did for the higher fundamentals. But at S/N ratios smaller than this, jnds increased much more sharply with further decreases in S/N ratio. This behaviour, combined with the larger jnds found for the lower fundamentals in quiet, meant that even changes of 10% in

fundamental frequency were not noticeable when the S/N ratio was worse than 10 dB. In short, when the perception of fundamental frequency is based on resolvable harmonics, it is relatively resistant to noise; when it is based on the temporal pattern of interacting harmonics, noise, once it reaches a certain level, has a much greater effect.

Assuming that this difference across fundamental frequencies is indeed related to the extent to which the harmonics are resolved, we can extrapolate to what would be found for 1/3-octave filters centred at other frequencies. Clearly, the limiting fundamental at which performance was similar to that for a pure tone would increase (as does auditory filter bandwidth) with increasing frequency. Given that auditory filters are of roughly constant Q between 1 and 4 kHz, fundamental frequencies greater than about 100 or 400 Hz should be relatively resistant to noise when listening through auditory filters centred at 1 or 4 kHz respectively. For speech, this means that for the frequency region at 1 kHz and below, nearly all voiced utterances will have their harmonics resolved, and hence the perception of voice pitch will be relatively resistant to noise. In the frequency region near 4 kHz and above, however (where there is little voiced energy anyway), the harmonics of most voiced speech will interact within a single auditory filter, and the perception of voice pitch will be relatively susceptible to noise masking. (This implies that, for a given signal-to-noise ratio in mid-frequency regions, it should be easier, on average, to hear the voice pitch of a woman.)

In summary, the information about voice pitch in low-frequency regions is perceptually more important than that in high-frequency ones for two reasons. Firstly, there is more energy there, and secondly, the low-frequency auditory filters are narrower. Although information about voice pitch is spread up to frequencies of about 4 kHz, both the acoustic structure of speech and the properties of auditory filters make it likely that only the lower frequency region (up to 1 or 2 kHz) is of perceptual importance, in quiet *and* in noise.

ii Performance of hearing-impaired listeners

Anyone who deals regularly with patients using hearing aids cannot help but be impressed by how often they complain about the difficulties they experience hearing speech in noise. It seems likely that some part of their problems may arise from a degradation in frequency selectivity. Given the advantages of a frequency-selective system in overcoming some of the effects of competing sounds, we should expect a broadening of auditory filters to make hearing-impaired listeners more sensitive to noise, even when making discriminations that do not themselves rely heavily on frequency selectivity. For example, a narrow band of competing noise, isolated in its effects by normal auditory filtering, might appear across a wide range of auditory filters in the impaired

ear, decreasing the signal-to-noise ratio through them all, and hence interfering with temporal judgements, such as detecting the presence of a silent gap indicating consonantal closure.

Most studies of the relationship between reduced frequency selectivity and speech perception in noise have taken a correlational approach. In these, a number of speech and psychoacoustical tests are applied to a set of hearing-impaired listeners, and some statistical measures of the relationships among the results of the various tests calculated (via factor analyses and/or simple correlations). The psychoacoustical biases of the experimenters involved in this work are readily apparent. Usually only the grossest measures of speech perception are used, with no attempt to relate the extensive psychoacoustical results to the acoustic structure of the speech sounds used. Typically, these incorporate such a wide range of acoustic features that such a task would be fruitless anyway. Stelmachowicz *et al.* (1985), for example, determined psychoacoustic tuning curves at 2 kHz and extracted five distinct summary measures from the results. These were correlated with 'speech perception ability' in broadband and low-pass noise, measured by determining the signal-to-noise ratio necessary to attain 75% correct in the N.U. Auditory Test No. 6, a set of lists of consonant-vowel-consonant words. It surely cannot be right that one (albeit important) auditory ability at one particular frequency needs five parameters to describe it, but that 'speech perception' in noise of words encompassing the total range of voicing, place and manner contrasts possible in English needs only one.

Even so, such studies have given some interesting results. There is usually a significant correlation between measures of frequency selectivity and speech perception in noise (Dreschler and Plomp, 1980, 1985; Festen and Plomp, 1983; Horst, 1982; Lyregaard, 1982; Stelmachowicz *et al.*, 1985; Tyler *et al.*, 1982) but not between selectivity and speech perception in quiet (Dreschler and Plomp, 1980, 1985; Festen and Plomp, 1983).

Dreschler (1983) and Dreschler and Plomp (1985) have extended the range of this type of study in two important ways. Firstly, the auditory capabilities of the hearing-impaired listeners were characterized at three frequencies, 0.5, 1.0 and 2.0 kHz. Secondly, they have attempted to relate auditory abilities not only to gross measures of speech perception, but also to specific perceptual features. Listeners were required to identify a set of nonsense consonant-vowel-consonant utterances in quiet and in noise. Multidimensional scaling of the confusion matrices using a technique known as INDSCAL (Carrol and Chang, 1970) allows the determination both of a map of the perceptual relationship between the sounds, and of the extent to which each of the listeners differentially weights the extracted perceptual dimensions. Unfortunately, the reliability of most of the parameters extracted from performance in noise was very low, and so it was only possible to analyse such results from initial consonants. Interestingly, no measure related to

frequency selectivity* correlated significantly with the use of features involved in identifying initial consonants in noise. For sounds in quiet, there was a tendency for relatively large increases in auditory filter bandwidth with frequency to be correlated with heavy weighting of the low-frequency region of speech (F_1 in vowels, and nasality/voicing for consonants).

Clearly, more studies along these lines are necessary, but it may prove more fruitful to analyse the results with respect to what is known (or measured) about the acoustic structure of the speech sounds used, rather than to rely on multidimensional scaling techniques to extract the relevant features. (Synthetic stimuli could also play an important role here, especially when care is taken to ensure their naturalness, see Fourcin, 1980; Hazan and Fourcin, 1985.) Although multidimensional scaling has the advantage that it requires no *a priori* statement of the relevant perceptual dimensions, there is no guarantee that the derived structure can be sensibly interpreted, given the effects of measurement noise and the desirability of solutions of low dimensionality. In Dreschler and Plomp's (1985) study, for example, the vowels seem well described by two readily interpretable dimensions, whereas the consonants are not. Although dimension I of the initial consonants is clearly a 'voicing' type dimension (with some oddities), dimension II is a hodgepodge, with /s/ and /z/ having extreme values, but sounds like /m/, /n/, /f/ and /k/ being nearly identical. More carefully constructed speech tests, with explicitly defined acoustic parameters, are likely to give results more clearly related to psychoacoustical abilities.

Finally, it should be noted that it will never be possible to account completely for speech perceptual ability simply through a consideration of sensory factors, no matter how thorough. As we discussed at the beginning of this chapter, speech understanding is a highly complex cognitive process, and these central factors have a large influence on overall performance, especially when the listeners' task becomes more similar to natural language processing.

4 COMPENSATING FOR IMPAIRED FREQUENCY SELECTIVITY

When hearing impairment was considered to be primarily a loss of sensitivity (as it often was before conductive losses could be treated so successfully by surgery), it was only natural to attempt to compensate for it by frequency-weighted amplification. Now that it is apparent that detection is a necessary but not sufficient condition for the perception of speech, there have been

*Frequency selectivity, in this study, was determined in a critical ratio task, perhaps not the best measure, but one which has been shown to be related to frequency selectivity, e.g. Dreschler and Plomp, 1980; Festen and Plomp, 1983; see Chapter 3.

numerous attempts to develop methods of acoustic signal processing that will correct for the other deficiencies found in impaired hearing. Hence, the introduction of amplitude-compression hearing aids can be seen as one attempt to solve the problem of limited dynamic range, or recruitment. Similarly, there have been a number of investigations of ways in which one might, through processing of the speech signal, compensate for degraded frequency selectivity.

A *Formant separation by ear*

One of the first methods followed naturally from the work of Pickett and his colleagues, reviewed in Section 3Cvi, which showed the deleterious effects the presence of a first formant could have on the detectability of second formant transitions. Danaher and Pickett (1975) argued that if upward spread of masking (a peripheral phenomenon) were responsible for the effect of F_1 on F_2, then presenting the formants independently to each ear should improve the detectability of transitions of F_2 compared to when both formants were presented to the same ear. On average, they found that this was indeed the case for hearing-impaired listeners, especially when F_1 had an initial transition as well. There was much variability, however, with one listener showing no masking effect of F_1, even under monotic presentation, some showing a masking effect only monotically, but not with dichotic presentation, and some even showing a 'central-masking effect' where the presence of F_1 in the ear opposite to F_2 impaired performance. Presumably, some of these differences are related to the differences in frequency selectivity across listeners, but this was not directly measured.

Turek *et al.* (1980) extended these studies to a rather more speech-like task. Impaired listeners were asked to identify as /bɑ/, /dɑ/ or /gɑ/ a continuum of three-formant synthetic stimuli in which place of articulation was cued only by the second and third formant transitions. Sometimes all the formants were presented to one ear, and sometimes F_1 was presented to one ear, while F_2 and F_3 were presented to the other. Again, there was great variability in listeners' responses. Although dichotic presentation of the formants did not generally improve performance, some listeners *were* greatly helped by this manipulation. Again, there was no attempt to assess frequency selectivity independently.

There is thus evidence that some listeners might benefit from a dichotic presentation of formants. Technically, this would be a difficult task, but there has been little exploration of even the more realistic possibility (given current technology) suggested by Danaher and Pickett of presenting low- and high-frequency bands, derived from simple linear filtering, to opposite ears. Franklin (1975, 1979, 1981) has advocated this approach, using as supporting evidence her findings that a low-frequency band of energy only increases consonantal recognition scores for the hearing impaired when it is presented

to the ear opposite the one receiving a high-frequency band of energy. Adding the low-frequency energy into the same ear has little, or even deleterious effects upon performance, depending on its level (Franklin, 1975, 1979). Unfortunately, although suggestive, her studies do not directly indicate that the division of speech into two dichotic bands would be advantageous, because the bands she used were artificially narrow (240–480 Hz and 1020–2040 Hz). Initial research by Kaplan and Pickett (1981) shows no advantage of presenting the speech band above 1 kHz to one ear, and the band below 1 kHz to the other over that of presenting the same signal to both ears. Haas (1982), too, found no significant differences between two such conditions, using a cut-off frequency of 780 Hz. Clearly, this question deserves further laboratory and clinical investigation, with particular emphasis on identifying patients most likely to benefit from such a scheme, and their having the opportunity to get used to the processing through the use of a wearable device.

B *Spectral pattern sharpening*

Another processing scheme which has received some attention aims to deal more generally with the 'blurred' representation of spectral patterns expected from degraded auditory filtering. Its rationale is based on a simple linear excitation-pattern model (recognized to be clearly inadequate) in which the representation of the 'inner spectrum' or excitation pattern (as seen, perhaps, on the basilar membrane or auditory nerve) is the result of the convolution, on appropriate scales (e.g. in Barks or ERBs), of the physical spectrum and the amplitude response of the auditory filter (see Chapter 5). In such a model, the loss of spectral detail resulting from convolution with a widened auditory filter can, within limits, be compensated for by a physical spectrum with enhanced detail.

When dealing with synthetic speech from a formant synthesizer, this is most easily effected by narrowing the formant bandwidths. Figure 7.31 gives some examples of the spectral envelope of a vowel that has had its formant bandwidths manipulated to be both wider and narrower than normal.

The first explicit test of such spectral enhancement was by Boers (1980) who digitally processed a set of natural Dutch sentences in such a way as to, among other things, increase the level differences between peaks and valleys in the spectrum. Their intelligibility before and after processing was assessed by determining the speech reception threshold (SRT) in speech-shaped noise; that is, the speech-to-noise ratio at which the sentences were just intelligible (Plomp and Mimpen, 1979).

On the whole, processing reduced intelligibility, uniformly so for the normal-hearing listeners. Unfortunately, Boers did not test a condition in which the signal underwent all the processing steps except the one meant to enhance spectral detail (e.g. filtering and digitization), so it is not known whether

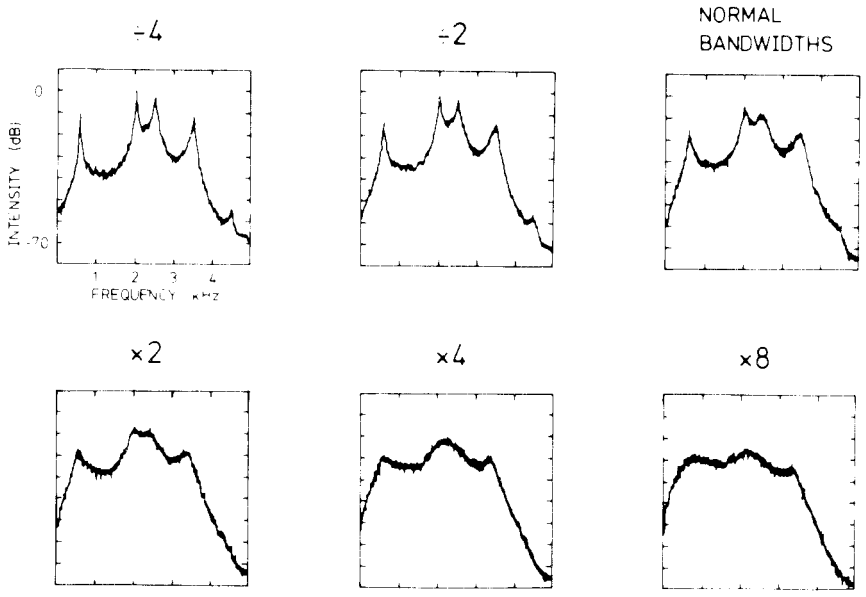


FIG. 7.31 Long-term power spectra of a synthesized vowel /ɛ/ showing the effects of varying formant bandwidths from 0.25 to 8 times their nominal normal values. The five resonators (representing five formants) were noise excited making the resulting stimuli sound as if they were whispered. This was done 'to define the complete spectral envelope and to avoid uncontrolled coincidences of harmonics with formant peaks.' Reproduced with permission from Summerfield *et al.* (1985).

this reduction in performance was due to the changes in the spectrum, or undesirable side effects of the processing (e.g. the restriction of signal bandwidth to 300–4800 Hz). Interestingly, two of the impaired listeners (of six) did show some small improvement (0.5–1.5 dB) with the processed signals, but it was not determined whether the frequency selectivity of these two listeners was different from that of the others.

Summerfield *et al.* (1985) investigated the same issue using a rather simpler set of stimuli—'whispered' (noise-excited) synthetic exemplars of 'bet', 'debt', 'get' or 'bib', 'bid', 'big' in which the phonemic contrasts were cued only by the second and third formant transitions. Formant bandwidths ranged from 0.25 to 8 times normal (as seen in Fig. 7.31). In addition, the frequency selectivity of all the listeners was determined with a three-point measure of the psychoacoustic tuning curve (PTC) in the region where the formant transitions occurred. For both normal and impaired listeners, making the formant bandwidths wider than normal led to worse performance. Making them narrower only affected the identification of final consonants, with a tendency for performance to be best when bandwidths were half their nominal normal value. There was only one systematic correlation between overall performance

and a measure of selectivity once the variability due to age and absolute threshold was partialled out—that between the high-frequency slope of the PTC and discriminability for the final set of consonants.

It is difficult to account for these results with an excitation-pattern model like the one discussed in Section IIIBiv. If we assume that some aspect of this task is related to the discrimination of static spectral patterns, then performance should increase with decreasing bandwidth until the slope of the spectral pattern is sharper than that of the auditory filters, at which point it should increase no more. This seems consistent with the results from normal listeners, but then the performance of impaired listeners (with widened filters) should level off at a bandwidth wider than that found for the normals. No indication of this is found, especially for the final consonants, where best performance is found at the same bandwidth for both normals and the hearing impaired. It may be that temporal aspects of the outputs of the auditory filters are more important in determining the percepts of the hearing impaired than of normal listeners. Summerfield *et al.* speculate that other psychoacoustical factors (e.g. the degree of backward masking) may be dominant in speech perceptual disabilities, especially for initial consonants.

If such temporal factors are crucial, then perhaps impaired listeners would show greater improvements with formant narrowing for relatively steady-state sounds. Sidwell and Summerfield (1985) have demonstrated that the peak-to-valley differences in the 'internal spectrum' of a sound with a vowel-like spectral envelope (as determined in masking experiments) can be enhanced by a process similar to formant narrowing. Cockitt and Pick (1986, personal communication), Jamieson *et al.* (1985) and Peters and Watkins (1984, personal communication) have all tested impaired listeners' abilities to identify vowels with manipulated formant bandwidths. In each case, some listeners have benefitted from the reduction of bandwidths, but it is not yet clear why other listeners (the majority) do not.

One way to obtain more comprehensible data may be to restrict the range of disabilities shown by patients. For instance, Summerfield *et al.* have set a good example for others in this area by explicitly measuring the frequency selectivity of the listeners they used. However, their interpretation of some of the results seems doubtful. Often, a masker that is remote in frequency from the probe has a larger masking effect than one at the probe frequency, which Summerfield *et al.* interpret as indicating very poor tuning. Some such cases may reflect the detection of beats (see Chapter 6 for a discussion of W-shaped tuning curves) although the tonal masker and narrowband noise signal used should minimize these. But there is another possibility. If the power spectrum model of masking is to be taken seriously, a masker remote in frequency from a probe can only have a larger masking effect than a masker at the probe frequency if the listener is operating through an auditory filter which is most sensitive to frequencies remote from the probe. In other words,

detection of energy at 2 kHz may be mediated through an auditory filter centred at 1 kHz, and hence a masker at 0.75 kHz may be a more effective masker of a probe at 2 kHz than a masker at 2 kHz. It is not so much that there is a poorly tuned auditory filter at 2 kHz, but that there is no auditory filter at all. This interpretation seems most plausible for the three impaired listeners who show more remote masking from the low-frequency side only; for maskers above the signal in frequency, masking decreases as the masker frequency increases. It seems likely that the strategies for speech perception used by such listeners, and those with some, even poor, genuine tuning, will differ considerably.

C *Spectral transposition*

Another possible way to minimize the masking effects of formants on one another, and to increase spectral detail, is to move the formants farther apart in frequency. Of course, this means some information must be lost, but the discarded parts of the signal (say the F_3 region and above for voiced speech) may well be unusable by an impaired listener. The loss of fricative information this entails could be dealt with in other ways, for instance by mapping only the highest frequencies down (Velmans, 1973, 1974; Velmans and Marcuson, 1983) or synthesizing low-frequency noises when high-frequency frication is detected (Johansson and Lindblad, 1971; Fourcin *et al.*, 1979). Ironically, nearly all the proposed spectral transposition techniques involve a *compression* of the spectrum into the reduced area of residual hearing (Braida *et al.*, 1979), a manipulation which will make even more severe demands on an auditory filter bank. Such processing seems of limited value even in normal listeners (e.g. Reed *et al.*, 1983) and will almost certainly prove disastrous with the hearing-impaired.

Some initial explorations of the possibilities of spectral expansion schemes (combined with bandwidth narrowing) have been made in our laboratory (Rhodes, 1984). One young hearing-impaired listener with an auditory filter bandwidth at 1 kHz about four times normal was tested on a synthetic /bɑ/–/dɑ/ continuum. One- and two-formant syntheses were used, with the second formant in the vowel always at 1234 Hz and the place contrast cued by varying the second formant transition. Normally, F_1 had a transition rising to a steady state of 765 Hz, but it could also rise only to 500 Hz, which was termed the 'stretched' condition. Sometimes, only F_2 was presented, in order to assess the extent to which F_1 interfered with the perception of the contrastive F_2 transition. Finally, three bandwidths were possible for the two formants: 'normal', 'wide' (normal $\times 2$) and 'narrow' (normal $\div 2$). The slope of the categorization function, as determined by a maximum-likelihood fit of a cumulative Gaussian curve (Bock and Jones, 1968), was used as the index of performance: sharper slopes indicate better performance.

For the standard frequency values of F_1 and F_2, there was a tendency for better performance with narrower bandwidths, although this did not reach statistical significance. In general agreement with Summerfield *et al.*'s (1985) findings, appropriate categorization of the wide-formant stimuli was not possible. In agreement with Hannley and Dorman's (1983) results for F_1 attenuation, best performance by far was for F_2 on its own, said to 'sound like Donald Duck'. Next best categorization was found with narrow formants when F_1 was lowered (the 'stretched' condition). Categorization of 'stretched' normal bandwidth formants was a little worse yet, but still better than that found with normal or narrow formants at the standard frequency values.

In summary, it appears that the effects of 'stretching' formants apart are more important than those of narrowing them, although some advantage might be gained by combining both techniques. One important problem that remains to be solved is to determine how formants can be moved about without losing the appropriate phonetic identity. In the experiments described above, for instance, although lowering F_1 helped in the identification of /b/ or /d/, it of course led to a completely different vowel quality.

It could be argued that since the *contrasts* among sounds would still be preserved, users would learn to use the information appropriately. If this turned out not to be true, it would be necessary to try to maintain, as far as is possible, the normal quality. Sidwell and Summerfield (1985) propose, for example, that F_1 could remain where it was but that F_2 and the higher formants could be replaced with a single formant at F_2', the position that leads to the best match for a full-formant vowel (see Section 3Bviii). This has three conceivable drawbacks. Firstly, for vowels in which F_3 and the higher formants are distant from F_2, F_2' is nearly identical with F_2, so no expansion would be effected. Secondly, when F_2' *is* different from F_2, it is always higher in frequency, and so more likely to fall in a region of greater loss for impaired listeners. Finally, the use of F_2' has substantial limitations when the F_2 and F_3 consonant-vowel formant transitions are in different directions, since averaging may cancel the transition. It may well be that for listeners with extreme impairments, presentation of a single formant parameter, mapped into the region of best auditory abilities, may be of most use. Even if extensive relearning is necessary, the basic patterns will still be naturally derived. A similar technique has already been applied in speech processing for multichannel cochlear implants (Tong *et al.*, 1980), an example of the speech pattern approach advocated by the EPI Group (Fourcin *et al.*, 1979, 1984).

D *Fundamental frequency pattern simplification and mapping*

The philosophy of hearing aid design underlying this technique is rather different to that almost universally held otherwise, and so is perhaps worthy of some amplification. Its basic premise is that an impaired listener will be

most helped by an explicit attempt to determine those speech patterns that are most needed, and to match them optimally to his residual sensory abilities. First developed in the context of single-channel extracochlear electrical stimulation for the totally deaf (Fourcin *et al.*, 1979, 1984), it was argued that since complete speech understanding by auditory means alone seemed unlikely via a single electrode (which is certainly true for most users of single-channel systems despite the claims of Hochmair-Desoyer *et al.*, 1985), it would be best to concentrate on the speech patterns most useful to lipreading. The fundamental frequency of the speaker's voice is ideal for this, as the information it carries is relatively invisible, and its frequency and amplitude characteristics are well matched to the psychophysical properties of single-channel stimulation. It is important to remember, though, that the approach is not limited to fundamental frequency. A sensory channel with sufficient capacity could make use of more complex input signals; the essential feature is that the desired speech patterns are explicitly extracted and their physical form matched to the patients' abilities.

The same processing used for electrically-stimulated patients is also being used for patients with severe hearing impairments. For the latter, the fundamental frequency of speech is extracted and presented as an acoustic sinusoid, which we have found to lead to best discrimination of frequency changes. Only one patient has yet been equipped with a take-home device of this sort (listener C described in Section 3Av) but our more recent work indicates that many listeners would be helped by such a Sinusoidal Voice (SiVo) aid. The reason why such an extreme simplification of speech is to the benefit of these patients (more so than presentation of the entire speech signal) is two-fold (Rosen and Fourcin, 1983). Firstly, they are unable to use the discarded information. C, for instance, shows almost no ability to use spectral pattern information, even in relatively steady-state vowel contrasts. Secondly, the information they can use, that relating to voice pitch, is much more easily perceived when presented as a single sinusoid (see Section 3Av for why this might be so).

The speech pattern approach has another advantage. Once the desired patterns (like fundamental frequency) are extracted, they need not be presented with the same absolute frequency values at which they occurred. They can be expanded and mapped into a region where patients have better sensitivity, while still preserving their essential pattern features. In the case of voice pitch, both for the electrically-stimulated and SiVo patients, it has proved useful to map frequencies of high-pitched speakers downwards by 50 Hz (some results of which are described in Section 3Av), since frequency discrimination worsens with increasing frequency. Again, this manipulation can be applied to any speech pattern (e.g. a single formant, as described in the previous section), although there may be limits in the extent to which patterns can be mapped before extensive relearning is needed to use them.

E *Choosing appropriate frequency emphasis*

Most of the techniques we have described so far are rather esoteric, and, apart from prototype work with SiVo aids, none are yet available in a commercial hearing aid. This should not be taken to mean that the frequency selectivity of a particular patient need only be taken into account when such aids become available. It seems likely that even the standard process of fitting hearing aids by manipulation of their frequency response could be improved by some knowledge of the auditory filtering properties of a particular patient.

Tyler *et al.* (1984) have already speculated on this, in light of their findings that auditory filters in the hearing impaired can be highly asymmetric, in either direction. Listeners with medium- to high-frequency filters that have shallow low-frequency slopes, but reasonably steep high-frequency ones, may benefit more from hearing aids with a frequency response that rises with frequency. This would tend to minimize the masking effects of F_1 on the higher formants. If this direction of asymmetry is most frequent among the impaired, as Tyler *et al.* found, it may account for the fact that rising frequency responses are generally thought to be preferable. On the other hand, listeners with filters having shallow high-frequency slopes, but reasonably steep low-frequency ones, would probably do better with more low-frequency amplification.

Measures of frequency selectivity also make it possible to determine when detection thresholds are mediated through auditory filters remote from the particular frequency in question (a possibility already considered in Section 4B). For example, although an audiogram may show that a patient is able to detect energy at 1 kHz, this may be 'heard' through an auditory filter most sensitive to 700 Hz or so. It seems to us, from the results of some preliminary studies, that many patients with so-called 'left-hand corner' audiograms (i.e. measurable thresholds only between 125 Hz and 1 kHz, with loss increasing with frequency) have auditory filters with centre frequencies that extend only up to 300 Hz or so. Clearly, the frequency response of an aid appropriate for such patients will be quite different to that appropriate for a patient who has auditory filters throughout a wider frequency range, even if degraded.

F *Increasing the signal-to-noise ratio*

Most of the schemes we have discussed for alleviating some of the perceptual difficulties caused by degraded frequency selectivity deal only with the structure of the speech signal itself. Yet it seems likely that the biggest handicap imposed by broad auditory filtering is the greatly increased disruption of speech understanding by extraneous sounds, simply because these affect the perception of *all* speech contrasts, whether primarily dependent on frequency selectivity or not. It is, therefore, certainly the case (as

Summerfield *et al.* (1985) also note) that processing schemes which isolate speech from background noise (whether environmental or from other speakers) will turn out to be at least as important as alterations in the structure of speech itself. One very attractive feature of speech pattern techniques is that they assign the problem of noise rejection to the pattern extraction circuits. Although current processing algorithms are usually easily disrupted by noise (especially those able to be implemented in portable battery-powered form), there is the likelihood of great progress in noise-resistant techniques, many of which may be fruitfully based on the processing employed by normal listeners.

5 FINAL COMMENTS

Although there are considerable lacunae in our knowledge, it is clear that many aspects of the processes of speech perception are illuminated by a consideration of the frequency selective properties of the auditory system. Our understanding of this interaction is likely to be furthered in two ways: empirically, from the increasing use of stimuli more closely related to speech in psychoacoustical testing, and theoretically, from thorough investigations of models which integrate time with place information in accounting for auditory percepts.

These developments will also influence clinical practice for impaired listeners, mostly via special-purpose signal-processing hearing aids. Given the rapid progress in electronics technology, it is likely that the development of such aids will be hindered more by conceptual than technical difficulties. New clinical techniques for identifying the appropriate aid for a patient, and fitting it adequately, will need to be developed. These will almost certainly include measures of frequency selectivity, even for relatively orthodox aids.

In our opinion, studies aimed at understanding the difficulties of the impaired listener will make faster progress if two principles are adhered to. Firstly, intensive investigation of a small number of listeners is likely to be more informative than cursory testing of a large number. This is desirable both to ensure the stability of results (which can be notoriously variable from session to session in hearing-impaired listeners) and to determine listeners' abilities in a sufficiently wide range of auditory tests. At the least, some basic measures of resolution in intensity, frequency and time are necessary. Secondly, the speech perceptual abilities of the listeners must be tested and discussed in an analytic way which focusses on the perception of specific acoustic pattern features. Two main techniques allow this: natural speech may be modified to neutralize some cues to a particular phonetic contrast while retaining others (e.g. Revoile *et al.*, 1982); similarly, the use of synthetic speech gives complete control over the acoustic patterns used (e.g. Hazan and

Fourcin, 1985). Of crucial importance to the latter is the use of reasonably natural sounding stimuli which allow the possibility of testing with at least the major acoustic cues to a particular contrast (Fourcin, 1980).

Finally, a more complete understanding of the relationship between speech perception and frequency selectivity, in both hearing-impaired and normal listeners, will surely require as much consideration of the detailed structure of speech, as of the properties of auditory filters. This explicit combination of speech and auditory science will illuminate both areas, and also provide a new range of tools for effective and practical approaches to training and prosthetic design.

6 SUGGESTIONS FOR FURTHER READING

The topics we have covered in this chapter range over a wide variety of disciplines, from acoustics and electrical engineering to psychology and phonetics. For those readers wishing to further their knowledge, the following sources are recommended.

Ladefoged (1975) provides an introductory course in classical linguistically oriented phonetics, though with a good introduction to spectrographic analysis. Fant (1960) and Flanagan (1972) both give thorough developments of the acoustic theory of speech production, but suitable perhaps only for those with sufficient prior training in physics or engineering. Pickett's (1980) book reviews most aspects of the acoustic structure of speech, in a less technical exposition. It also includes extensive discussion of many issues in speech perception. Lehiste's (1967) collection, as well as that of Fry (1976), contain many of the most influential papers regarding the acoustic structure of speech and its perception in single convenient volumes. Chapter 6 in Plomp (1976) contains a particularly good discussion of the relationship between frequency selectivity and the perceived timbre of complex tones, which has heavily influenced parts of the discussion in Section 3B.

7 ACKNOWLEDGEMENTS

Grateful appreciation is extended to all those who permitted figures to be reproduced here: S. P. Bacon, R. A. W. Bladon, R. Carlson, M. T. Hannley, M. Hirano, A. Hoekstra, R. Plomp, L. Pols, R. Shepard and Q. Summerfield.

Further thanks to: Nick Noscoe of the Middlesex Hospital for the xeroradiograms of Figs 7.1 and 7.3; Virginia Ball for her help in making spectrograms; to Virginia Ball, Rachel Pearn, Michael Shailer, Brian Glasberg

and Brian Moore for assistance in proof-reading; to Evelyn Abberton for critical reading of the text and discussions on phonetic issues; and especially, to the editor, Brian Moore, not only for his numerous suggestions for improving this chapter, but also for his patience.

The first author is supported by the Medical Research Council of the United Kingdom.

REFERENCES

Abberton, E., Fourcin, A. J., Rosen, S., Walliker, J. R., Howard, D. M., Moore, B. C. J., Douek, E. E. and Frampton, S. (1985). Speech perceptual and productive rehabilitation in electro-cochlear stimulation. In *Cochlear Implants* (eds R. A. Schindler and M. M. Merzenich), Raven Press, New York.

Ananthapadmanabha, T. V. and Fant, G. (1982). Calculation of true glottal flow and its components. *Speech Trans. Lab.—Q. Prog. Stat. Rep.* **1**, 1–30, (Royal Institute of Technology, Stockholm).

Bacon, S. P. (1979). Suppression effects in vowel pulsation patterns. M.A. thesis, University of Kansas.

Bacon, S. P. and Brandt, J. F. (1982). Auditory processing of vowels by normal-hearing and hearing-impaired listeners. *J. Speech Hear. Res.* **25**, 339–347.

Bailey, P. J. and Summerfield, Q. (1980). Information in speech: observations on the perception of [s]-stop clusters. *J. Exp. Psychol.: Hum. Percept. Perf.* **6**, 536–563.

van den Berg, J. W. (1958). Myoelastic-aerodynamic theory of voice production. *J. Speech Hear. Res.* **1**, 227–444.

van den Berg, J. W. (1968). Mechanism of the larynx and the laryngeal vibrations. In *Manual of Phonetics* (ed. B. Malmberg), North Holland, Amsterdam.

Bladon, A. (1984). Rapid versus slow spectral change: implications for dynamic auditory processing of speech. *Proc. Inst. Acoust.* **6**, 275–280.

Bladon, R. A. W. and Lindblom, B. (1981). Modeling the judgment of vowel quality differences. *J. Acoust. Soc. Am.* **69**, 1414–1422.

Bock, R. D. and Jones, L. V. (1968). *The Measurement and Prediction of Judgment and Choice*. Holden-Day, San Francisco.

Boers, P. M. (1980). Formant enhancement of speech for listeners with sensorineural hearing loss. *IPO Annual Progress Report* 15, 21–28, Institute for Perception Research, Eindhoven.

Braida, L. D., Durlach, N. I., Lippmann, R. P., Hicks, B. L., Rabinowitz, W. M. and Reed, C. M. (1979). *Hearing aids—A review of past research on linear amplification, amplitude compression and frequency lowering*. ASHA Monographs No. 19, American Speech-Language-Hearing Association.

Bregman, A. S. and Pinker, S. (1978). Auditory streaming and the building of timbre. *Can. J. Psychol.* **32**, 19–31.

Buus, S. (1985). Release from masking caused by envelope fluctuations. *J. Acoust. Soc. Am.* **78**, 1958–1965.

Carlson, R. and Granström, B. (1979). Model predictions of vowel dissimilarity. *Speech Trans. Lab. — Q. Prog. Stat. Rep.* **3–4**, 84–104, (Royal Institute of Technology, Stockholm).

Carlson, R., Granström, B. and Fant, G. (1970). Some studies concerning perception of isolated vowels. *Speech Trans. Lab. — Q. Prog. Stat. Rep.* **2–3**, 19–35, (Royal Institute of Technology, Stockholm).

Carlson, R., Granström, B. and Klatt, D. (1979). Vowel perception: the relative perceptual salience of selected acoustic manipulations. *Speech Trans. Lab. — Q. Prog. Stat. Rep.* **3–4**, 73–83, (Royal Institute of Technology, Stockholm.)

Carroll, J. D. and Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika* **35**, 238–319.

Ching, Y. C. (1981). Communication of lexical tone patterns in Cantonese. Ph.D. thesis, University College London.

Chistovich, L. A. (1985). Central auditory processing of peripheral vowel spectra. *J. Acoust. Soc. Am.* **77**, 789–805.

Chistovich, L. A., Granstrem, M. P., Kozhevnikov, V. A., Lesogor, L. W., Shupljakov, V. S., Taljasin, P. A. and Tjulkov, W. A. (1974). A functional model of signal processing in the peripheral auditory system. *Acustica* **31**, 349–353.

Chistovich, L. A., Sheikin, R. L. and Lublinskaja, V. V. (1979). "Centres of gravity" and spectral peaks as the determinants of vowel quality. In *Frontiers of Speech Communication Research* (eds B. Lindblom and S. Öhman), Academic Press, London, New York.

Cockitt, E. C. and Pick, G. F. (1986). The perception of spectrally degraded and spectrally enhanced vowels. *Br. J. Audiol.* (abstract) (in press).

Danaher, E. M. and Pickett, J. M. (1975). Some masking effects produced by low-frequency vowel formants in persons with sensorineural hearing loss. *J. Speech Hear. Res.* **18**, 261–271.

Danaher, E. M., Osberger, M. J. and Pickett, J. M. (1973). Discrimination of formant frequency transitions in synthetic vowels. *J. Speech Hear. Res.* **16**, 439–451.

Darwin, C. J. (1984). Perceiving vowels in the presence of another sound: constraints on formant perception. *J. Acoust. Soc. Am.* **76**, 1636–1649.

Darwin, C. J. and Gardner, R. B. (1986). Mistuning a harmonic of a vowel: grouping and phase effects on vowel quality. *J. Acoust. Soc. Am.* **79**, 838–845.

Delattre, P. C., Liberman, A. M., Cooper, F. S. and Gerstman, L. J. (1952). An experimental study of the acoustic determinants of vowel color; observations on one- and two-formant vowels synthesised from spectrographic patterns. *Word* **8**, 195–210. Also in *Acoustic Phonetics* (1976) (ed. D. B. Fry), Cambridge University Press, Cambridge.

Delattre, P. C., Liberman, A. M. and Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Am.* **27**, 769–773. Also in *Readings in Acoustic Phonetics* (1967) (ed. I. Lehiste), MIT Press, Cambridge, Mass. And in *Acoustic Phonetics* (1976) (ed. D. B. Fry), Cambridge University Press, Cambridge.

Delgutte, B. (1982). Some correlates of phonetic distinctions at the level of the auditory

nerve. In *The Representation of Speech in the Peripheral Auditory System* (eds R. Carlson and B. Granström), Elsevier Biomedical Press, Amsterdam.

Dorman, M. F., Dobbins, E., Hannley, M. T. and Lindholm, J. M. (unpublished manuscript). Factors underlying errors in identification of voiced stop consonants by hearing-impaired listeners.

Dorman, M. F., Raphael, L. J. and Isenberg, D. (1980). Acoustic cues for a fricative-affricate contrast in word-final position. *J. Phonetics* **8**, 397–405.

Dorman, M. F., Lindholm, J. M. and Hannley, M. T. (1985). Influence of the first formant on the recognition of voiced stop consonants by hearing-impaired listeners. *J. Speech Hear. Res.* **28**, 377–380.

Dreschler, W. A. (1983). Impaired frequency/time resolution and its effect on speech intelligibility. In *Hearing—Physiological Bases and Psychophysics* (eds R. Klinke and R. Hartmann), Springer-Verlag, Berlin.

Dreschler, W. A. and Plomp, R. (1980). Relation between psychophysical data and speech perception for hearing-impaired subjects. I. *J. Acoust. Soc. Am.* **68**, 1608–1615.

Dreschler, W. A. and Plomp, R. (1985). Relations between psychophysical data and speech perception for hearing-impaired subjects. II. *J. Acoust. Soc. Am.* **78**, 1261–1270.

Dunn, H. K. (1961). Methods of measuring formant bandwidths. *J. Acoust. Soc. Am.* **33**, 1737–1746.

Fant, G. M. (1956). On the predictability of formant levels and spectrum envelopes from formant frequencies. In *For Roman Jakobson* (eds M. Halle, H. Lunt and H. MacLean), Mouton, The Hague. Also in *Readings in Acoustic Phonetics* (1967) (ed. I. Lehiste), MIT Press, Cambridge, Mass.

Fant, G. M. (1960). *Acoustic Theory of Speech Production*, Mouton: 's-Gravenhage.

Fastl, H. (1976). Temporal masking effects: I. Broad band noise masker. *Acustica* **35**, 287–302.

Fastl, H. (1977a). Temporal masking effects: II. Critical band noise masker. *Acustica* **36**, 317–331.

Fastl, H. (1977b). Roughness and temporal masking patterns of sinusoidally amplitude modulated broadband noise. In *Psychophysics and Physiology of Hearing* (eds E. F. Evans and J. P. Wilson), Academic Press, London, New York.

Fastl, H. and Weinberger, M. (1981). Frequency discrimination for pure and complex tones. *Acustica* **49**, 77–78.

Festen, J. M. and Plomp, R. (1983). Relations between auditory functions in impaired hearing. *J. Acoust. Soc. Am.* **73**, 652–662.

Flanagan, J. L. (1955). A difference limen for vowel formant frequency. *J. Acoust. Soc. Am.* **27**, 613–617. Also in *Readings in Acoustic Phonetics* (1967) (ed. I. Lehiste), MIT Press, Cambridge, Mass.

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

Flanagan, J. L. and Saslow, M. G. (1958). Pitch discrimination for synthetic vowels. *J. Acoust. Soc. Am.* **30**, 435–442.

Fourcin, A. J. (1976). Speech pattern tests for deaf children. In *Disorders of Auditory Function II* (ed. S. D. G. Stephens), Academic Press, London and New York.

Fourcin, A. J. (1978). Acoustic patterns and speech acquisition. In *The Development of Communication* (eds N. Waterson and C. Snow), John Wiley, Chichester. Also in *Speech and Hearing: Work in Progress* (1978), 143–171, Phonetics and Linguistics, University College London.

Fourcin, A. J. (1980). Speech Pattern Audiometry. In *Auditory Investigation: The Scientific and Technological Basis* (ed. H. A. Beagley), Clarendon Press, Oxford.

Fourcin, A. J. (1981). Larynogographic assessment of phonatory function. In *Proceedings of the Conference on the Assessment of Vocal Pathology* (eds C. L. Ludlow and M. O'C. Hart), ASHA Reports 11, ASHA, Rockville, Maryland.

Fourcin, A. J., Evershed, S., Fisher, J., King, A., Parker, A. and Wright, R. (1978). Perception and production of speech patterns by hearing-impaired children. *Speech and Hearing: Work in Progress*, 173–204, Phonetics and Linguistics, University College London.

Fourcin, A. J., Rosen, S. M., Moore, B. C. J., Douek, E. E., Clarke, G. P., Dodson, H. and Bannister, L. H. (1979). External electrical stimulation of the cochlea: clinical, psychophysical, speech-perceptual and histological findings. *Br. J. Audiol.* **13**, 85–107.

Fourcin, A. J., Douek, E. E., Moore, B. C. J., Abberton, E., Rosen, S. and Walliker, J. (1984). Speech pattern element stimulation in electrical hearing. *Arch. Otolaryngol.* **110**, 145–153.

Franklin, B. (1975). The effect of combining low- and high-frequency passbands on consonant recognition in the hearing impaired. *J. Speech Hear. Res.* **18**, 719–727.

Franklin, B. (1979). A comparison of the effect on consonant discrimination of combining low- and high-frequency passbands in normal, congenital, and adventitious hearing-impaired subjects. *J. Am. Aud. Soc.* **5**, 168–176.

Franklin, B. (1981). Split-band amplification: a HI/LO hearing aid fitting. *Ear and Hearing* **2**, 230–233.

Fry, D. B. (1958). Experiments in the perception of stress. *Lang. Speech* **1**, 126–152. Also in *Acoustic Phonetics* (1976) (ed. D. B. Fry), Cambridge University Press, Cambridge.

Fry, D. B. (1976) (ed.). *Acoustic Phonetics*, Cambridge University Press, Cambridge.

Fry, D. B. (1968). Prosodic phenomena. In *Manual of Phonetics*, (ed. B. Malmberg), North Holland, Amsterdam.

Fujimura, O. (1962). Analysis of nasal consonants. *J. Acoust. Soc. Am.* **34**, 1865–1875. Also in *Readings in Acoustic Phonetics* (1967) (ed. I. Lehiste), MIT Press, Cambridge, Mass.

Fujimura, O. and Lindqvist, J. (1971). Sweep-tone measurements of vocal-tract characteristics. *J. Acoust. Soc. Am.* **49**, 541–558.

Gabrielsson, A., Johansson, B., Lindblad, A.-C., Persson, L., Pettersson, A. and Rosenqvist, B. (1975). Frequency discrimination for bands of noise. *Audiology* **14**, 1–20.

Gagné, J.-P. and Zurek, P. M. (1980). Frequency-difference limens for first formants of synthetic vowels. *Research Department Periodic Progress Report* No. 23, 9, Central Institute for the Deaf, St. Louis, Missouri.

Gerstman, L. J. (1957). Perceptual dimensions for the friction portions of certain speech sounds. Ph.D. thesis, New York University.

Gimson, A. C. (1962). *An Introduction to the Pronunciation of English*, Edward Arnold, London.

Goldstein, J. L. (1967). Auditory spectral filtering and monaural phase perception. *J. Acoust. Soc. Am.* **41**, 458–479.

Goldstein, J. L. and Srulovicz, P. (1977). Auditory-nerve spike intervals as an adequate basis for aural frequency measurement. In *Psychophysics and Physiology of Hearing* (eds E. F. Evans and J. P. Wilson), Academic Press, London, New York.

Haas, G. F. (1982). Impaired listeners' recognition of speech presented dichotically through high- and low-pass filters. *Audiology* **21**, 433–453.

Halle, M., Hughes, G. W. and Radley, J.-P. A. (1957). Acoustic properties of stop consonants. *J. Acoust. Soc. Am.* **29**, 107–116. Also in *Readings in Acoustic Phonetics* (1967) (ed. I. Lehiste), MIT Press, Cambridge, Mass., and *Acoustic Phonetics* (1976) (ed. D. B. Fry), Cambridge University Press, Cambridge.

Hannley, M. and Dorman, M. F. (1983). Susceptibility to intraspeech spread of masking in listeners with sensorineural hearing loss. *J. Acoust. Soc. Am.* **74**, 40–51.

Harris, J. D. (1952). Pitch discrimination. *J. Acoust. Soc. Am.* **24**, 750–755. Also in *Forty Germinal Papers in Human Hearing* (1969) (ed. J. D. Harris), *Journal of Auditory Research*, Groton, Connecticut.

Harris, K. S. (1958). Cues for the discrimination of American English fricatives in spoken syllables. *Lang. Speech* **1**, 1–7. Also in *Acoustic Phonetics* (1976) (ed. D. B. Fry), Cambridge University Press, Cambridge.

't Hart, J. (1981). Differential sensitivity to pitch distance, particularly in speech. *J. Acoust. Soc. Am.* **69**, 811–821.

Hawkins, J. E., Jr. and Stevens, S. S. (1950). The masking of pure tones and of speech by white noise. *J. Acoust. Soc. Am.* **22**, 6–13. Also in *Forty Germinal Papers in Human Hearing* (1969) (ed. J. D. Harris), *Journal of Auditory Research*, Groton, Connecticut.

Hazan, V. and Fourcin, A. J. (1985). Microprocessor-controlled speech pattern audiometry. *Audiology* **24**, 325–335.

Heinz, J. M. and Stevens, K. N. (1961). On the properties of voiceless fricative consonants. *J. Acoust. Soc. Am.* 33, 589–596. Also in *Readings in Acoustic Phonetics* (1967) (ed. I. Lehiste), MIT Press, Cambridge, Mass.

Henning, G. B. and Grosberg, S. L. (1968). Effect of harmonic components on frequency discrimination. *J. Acoust. Soc. Am.* **44**, 1386–1389.

Hirano, M. (1981). *Clinical Examination of Voice*, Springer-Verlag, New York.

Hochmair-Desoyer, I. J., Hochmair, E. S. and Stiglbrunner, H. K. (1985). Psychoacoustic temporal processing and speech understanding in cochlear implant patients. In *Cochlear Implants* (eds R. A. Schindler and M. M. Merzenich), Raven Press, New York.

Hoekstra, A. (1979). Frequency discrimination and frequency analysis in hearing. Ph.D. Thesis, University of Groningen.

Hoekstra, A. and Ritsma, R. J. (1977). Perceptive hearing loss and frequency

selectivity. In *Psychophysics and Physiology of Hearing* (eds E. F. Evans and J. P. Wilson), Academic Press, London and New York.

Holbrook, A. and Fairbanks, G. (1962). Dipthong formants and their movements. *J. Speech Hear. Res.* **5**, 38–58. Also in *Readings in Acoustic Phonetics* (1967) (ed. I. Lehiste), MIT Press, Cambridge, Mass.

Horst, J. W. (1982). Discrimination of complex signals in hearing. Ph.D. thesis, University of Groningen.

Horst, J. W. and Ritsma, R. J. (1982). Formant discrimination and speech intelligibility at low signal-to-noise ratios in subjects with a sensory-neural hearing loss. *J. Acoust. Soc. Am.* **72**, Suppl. 1, S90 (abstract).

House, W. F. (1976). Cochlear implants. *Ann. Otol. Rhinol. Laryngol.* **85**, Suppl. 27.

Houtgast, T. (1974). Auditory analysis of vowel-like sounds. *Acustica* **31**, 320–324.

Howell, P. (1981). Identification of vowels in and out of context. *J. Acoust. Soc. Am.* **70**, 1256–1260.

Jamieson, D. G., Ponton, W. and Espinoza-Varas, B. (1985). Reduction of formant bandwidth improves vowel identification with sensorineural impairment. *J. Acoust. Soc. Am.* **77**, Suppl. 1, S8 (abstract).

Johansson, B. and Lindblad, A. C. (1971). The use of compression and frequency transposition in hearing aids, *Scand. Audiology* **2**, 161–173.

Joos, M. (1948). Acoustic Phonetics. Language Monograph No. 23, *Language* **24** supplement.

Kaplan, H. and Pickett, J. M. (1981). Effects of dichotic/diotic versus monotic presentation on speech understanding in noise in elderly hearing-impaired listeners. *Ear Hear.* **2**, 202–207.

Kiang, N. Y. S. and Moxon, E. C. (1972). Physiological considerations in artificial stimulation of the inner ear. *Ann. Otol. Rhinol. Laryngol.* **81**, 714–730.

Klatt, D. H. (1973). Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception. *J. Acoust. Soc. Am.* **53**, 8–16.

Klein, W., Plomp, R., and Pols, L. C. W. (1970). Vowel spectra, vowel spaces and vowel identification. *J. Acoust. Soc. Am.* **48**, 999–1009.

Kuhl, P. K. (1976). Speech perception in early infancy: the acquisition of speech-sound categories. In *Hearing and Davis: Essays Honoring Hallowell Davis* (eds S. K. Hirsh, D. H. Eldredge, I. J. Hirsh and S. R. Silverman), Washington University Press, St. Louis, Missouri.

Koizumi, T., Shuji, T. and Sejiro, H. (1985). Glottal source–vocal tract interaction. *J. Acoust. Soc. Am.* **78**, 1541–1547.

Ladefoged, P. (1975). *A Course in Phonetics*, Harcourt Brace Jovanovich, New York.

Lehiste, I. (1967) (ed.). *Readings in Acoustic Phonetics*, MIT Press, Cambridge, Mass.

Liberman, A. M. (1982). On finding that speech is special. *Am. Psychol.* **37**, 148–167.

Liberman, A. M., Delattre, P. C., Cooper, F. S. and Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of stop and nasal consonants, Psychological Monographs, 68, No. 8. Also in *Acoustic Phonetics* (1976), (ed. D. B. Fry), Cambridge University Press, Cambridge.

Liberman, A. M., Delattre, P. C., Gerstman, L. J. and Cooper, F. S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *J. Exp. Psychol.* **52**, 127–137.

Liberman, A. M., Delattre, P. C. and Cooper, F. S. (1958). Some cues for the distinction between voiced and voiceless stops in initial position. *Lang. Speech* **1**, 153–167.

Liberman, A. M., Harris, K. S., Eimas, P., Lisker, L. and Bastian, J. (1961a). An effect of learning on speech perception: the discrimination of durations of silence with and without phonemic significance. *Lang. Speech* **4**, 175–195.

Liberman, A. M., Harris, K. S., Kinney, J. A. and Lane, H. (1961b). The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. *J. Exp. Psychol.* **61**, 379–388.

Licklider, J. C. R. (1957). Effects of changes in the phase pattern upon the sound of a 16-harmonic tone. *J. Acoust. Soc. Am.* **29**, 780 (abstract).

Licklider, J. C. R. and Miller, G. A. (1981). The perception of speech. In *Handbook of Experimental Psychology* (ed. S. S. Stevens), John Wiley, New York.

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

Lyregaard, P. E. (1982). Frequency selectivity and speech intelligibility in noise. In *Binaural Effects in Normal and Impaired Hearing* (eds O. J. Pedersen and T. Poulsen), *Scand. Audiol. Suppl.* 15.

Mack, M. and Blumstein, S. E. (1983). Further evidence of acoustic invariance in speech production: The stop-glide contrast. *J. Acoust. Soc. Am.* **73**, 1739–1750.

Mathes, R. C. and Miller, R. L. (1947). Phase effects in monaural perception. *J. Acoust. Soc. Am.* **19**, 780–797.

Mermelstein, P. (1978). Difference limens for formant frequencies of steady-state and consonant-bound vowels. *J. Acoust. Soc. Am.* **63**, 572–580.

Miller, G. A. and Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.* **27**, 338–352. Also in *Readings in Acoustic Phonetics* (1967) (ed. I. Lehiste), MIT Press, Cambridge, Mass., and *Acoustic Phonetics* (1976) (ed. D. B. Fry), Cambridge University Press, Cambridge.

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In *Perspectives on the Study of Speech* (eds P. D. Eimas and J. L. Miller), Lawrence Erlbaum, Hillsdale, New Jersey.

Miller, J. L. and Baer, T. (1983). Some effects of speaking rate on the production of /b/ and /w/. *J. Acoust. Soc. Am.* **73**, 1751–1755.

Miller, R. L. (1953). Auditory tests with synthetic vowels. *J. Acoust. Soc. Am.* **25**, 114–121.

Moore, B. C. J. (1983). Review paper: Psychoacoustics of normal and impaired listeners. In *Hearing—Physiological Bases and Psychophysics* (eds R. Klinke and R. Hartmann), Springer-Verlag, Berlin.

Moore, B. C. J. (1985). Additivity of simultaneous masking, revisited. *J. Acoust. Soc. Am.* **78**, 488–494.

Moore, B. C. J. and Glasberg, B. R. (1983a). Masking patterns for synthetic vowels in simultaneous and forward masking. *J. Acoust. Soc. Am.* **73**, 905–917.

Moore, B. C. J. and Glasberg, B. R. (1983b). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *J. Acoust. Soc. Am.* **74**, 750–753.

Moore, B. C. J., Glasberg, B. R. and Shailer, M. J. (1984). Frequency and intensity difference limens for harmonics within complex tones. *J. Acoust. Soc. Am.* **75**, 550–561.

Moore, B. C. J., Glasberg, B. R. and Peters, R. W. (1985). Relative dominance of individual partials in determining the pitch of complex tones. *J. Acoust. Soc. Am.* **77**, 1853–1860.

Nakagawa, T., Saito, S. and Yoshino, T. (1983). Discriminability of second formant frequencies in hearing-impaired children, *Ann. Bull. Res. Inst. Logoped. Phoniat.* **17**, 201–209, University of Tokyo.

Nierop, D. J. P. J. van, Pols, L. C. W. and Plomp, R. (1973). Frequency analysis of Dutch vowels from 25 female speakers, *Acustica* **29**, 110–118.

van Noorden, L. (1982). Two channel pitch perception. In *Music, Mind and Brain* (ed. M. Clynes), Plenum Press, New York.

Nord, L. (1976). Perceptual experiments with nasals. *Speech Trans. Lab. — Q. Prog. Stat. Rep.* **2–3**, 5–8, Royal Institute of Technology, Stockholm.

Noscoe, N. J., Fourcin, A. J., Brown, N. J. and Berry, R. J. (1983). Examination of vocal fold movement by ultra-short pulse X-radiography. *Br. J. Radiol.* **56**, 641–645.

O'Connor, J. D., Gerstman, L. J., Liberman, A. M., Delattre, P. C. and Cooper, F. S. (1957). Acoustic cues for the perception of initial /w,j,r,l/ in English. *Word* **13**, 24–43. Also in *Acoustic Phonetics* (1976) (ed. D. B. Fry), Cambridge University Press, Cambridge.

Owens, E., Talbott, C. B. and Schubert, E. D. (1968). Vowel discrimination of hearing-impaired listeners. *J. Speech Hear. Res.* **11**, 648–655.

Owens, E., Kessler, D. and Schubert, E. D. (1982). Interim assessment of candidates for cochlear implants. *Arch. Otolaryngol.* **108**, 478–483.

Pastore, R. E. (1981). Possible psychoacoustic factors in speech perception. In *Perspectives on the Study of Speech* (eds P. D. Eimas and J. L. Miller), Lawrence Erlbaum, Hillsdale, New Jersey.

Peters, C. J. and Watkins, A. J. (1984). Vowel perception and formant bandwidths. *Br. J. Audiol.* **18**, 250–251 (abstract).

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* **24**, 175–184. Also in *Readings in Acoustic Phonetics* (1967) (ed. I. Lehiste), MIT Press, Cambridge, Mass., and *Acoustic Phonetics* (1976) (ed. D. B. Fry), Cambridge University Press, Cambridge.

Pickett, J. M. (1980). *The Sounds of Speech Communication*, University Park Press, Baltimore.

Pickett, J. M. and Martony, J. (1970). Low-frequency vowel formant discrimination in hearing-impaired listeners. *J. Speech Hear. Res.* **13**, 347–359.

Pickett, J. M., Revoile, S. G. and Danaher, E. M. (1983). Speech cue measures of impaired hearing. In *Hearing Research and Theory*, Vol. 2, (eds J. V. Tobias and E. D. Schubert), Academic Press, New York and London.

Pierrehumbert, J. (1979). The perception of fundamental frequency declination. *J. Acoust. Soc. Am.* **66**, 363–368.

Pisoni, D. B., Carrell, T. D. and Simnick, S. S. (1979). Does a listener need to recover the dynamic vocal tract gestures of a talker to recognize his vowels? *Speech*

Communication Papers presented at the 97th Meeting of the Acoustical Society of America (eds J. J. Wolf and D. H. Klatt).

Plomp, R. (1970). Timbre as a multidimensional attribute of complex tones. In *Frequency Analysis and Perodicity Detection in Hearing* (eds R. Plomp and G. F. Smoorenburg), A. W. Sijthoff, Leiden.

Plomp, R. (1975). Auditory analysis and timbre perception. In *Auditory Analysis and Perception of Speech* (eds G. Fant and M. A. A. Tatham), Academic Press, London and New York.

Plomp, R. (1976). *Aspects of Tone Sensation*, Academic Press, London, New York.

Plomp, R. (1983). The role of modulation in hearing. In *Hearing—Physiological Bases and Psychophysics* (eds R. Klinke and R. Hartmann), Springer-Verlag, Berlin.

Plomp, R. and Mimpen, A. M. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology* **18**, 43–52.

Plomp, R. and Steeneken, H. J. M. (1969). Effect of phase on the timbre of complex tones. *J. Acoust. Soc. Am.* **46**, 409–421.

Plomp, R. and Steeneken, H. J. M. (1971). Pitch versus timbre. *Proc. 7th Int. Cong. Acoust.*, Budapest, **3**, 377–380.

Plomp, R. and Steeneken, H. J. M. (1973). Place dependence of timbre in reverberant sound fields. *Acustica* **28**, 50–58.

Plomp, R., Pols, L. C. W. and Geer, J. O. van der (1967). Dimensional analysis of vowel spectra. *J. Acoust. Soc. Am.* **41**, 707–712.

Pols, L. C. W. and Schouten, M. E. H. (1982). Perceptual relevance of coarticulation. In *The Representation of Speech in the Peripheral Auditory System* (eds R. Carlson and B. Granström), Elsevier Biomedical Press, Amsterdam.

Pols, L. C. W., Kamp, L. J. Th. van der, and Plomp, R. (1969). Perceptual and physical space of vowel sounds. *J. Acoust. Soc. Am.* **46**, 458–467.

Pols, L. C. W., Tromp, H. R. C. and Plomp, R. (1973). Frequency analysis of Dutch vowels from 50 male speakers. *J. Acoust. Soc. Am.* **53**, 1093–1101.

Reed, C. M., Hicks, B. L., Braida, L. D. and Durlach, N. I. (1983). Discrimination of speech processed by low-pass filtering and pitch-invariant frequency lowering. *J. Acoust. Soc. Am.* **74**, 409–419.

Repp, B. H. (1983). Trading relations among acoustic cues in speech perception are largely a result of phonetic categorization. *Speech Commun.* **2**, 341–361.

Revoile, S., Pickett, J. M., Holden, L. D. and Talkin, D. (1982). Acoustic cues to final stop voicing for impaired- and normal-hearing listeners. *J. Acoust. Soc. Am.* **72**, 1145–1154.

Rhodes, S. (1984). Improving perception of formant frequency transitions in hearing-impaired listeners. Undergraduate thesis, University College London.

Ricks, D. (1975). Vocal communication in pre-verbal normal and autistic children. In *Language, Cognitive Deficits and Retardation* (ed. N. O'Connor), Butterworth, London.

Risberg, A. (1974). The importance of prosodic speech elements for the lipreader. In *Visual and Audio-Visual Perception of Speech*, Sixth Danavox Symposium (eds H. B. Nielsen and E. Kampp), *Scandinavian Audiology*, Suppl. 4, Almquist and Wiksell, Stockholm.

Ritsma, R. J., Domburg, G. and Donders, J. J. H. (1967). Frequency discrimination of a single formant. *IPO Annual Progress Report* **2**, 30–36, Institute for Perception Research, Eindhoven.

Ritsma, R. J., Domburg, G. and Donders, J. J. H. (1968). On the response characteristic of the ear. *IPO Annual Progress Report* **3**, 7–13, Institute for Perception Research, Eindhoven.

Rosen, S. (1984). Hyperacute monaural phase sensitivity in the hearing-impaired. *Br. J. Audiol.* **18**, 257–258 (abstract).

Rosen, S. (1986a). Monaural phase sensitivity: frequency selectivity and temporal processes. In *Auditory Frequency Selectivity* (eds B. C. J. Moore and R. D. Patterson), Plenum, New York.

Rosen, S. (1986b). Phase and the hearing impaired. In *Proc. NATO-ARW 'The Psychophysics of Speech Perception'* (ed. M. E. H. Schouten), Martinus Nijhoff, The Netherlands (in press).

Rosen, S. and Ball, V. (1986). Speech perception with the Vienna extra-cochlear single-channel implant: a comparison of two approaches to speech coding. *Br. J. Audiol.* **20**, 61–83.

Rosen, S. and Fourcin, A. J. (1983). When less is more—further work. *Speech, Hearing and Language: Work in Progress* **1**, 3–27, Phonetics and Linguistics, University College London.

Rosen, S., Fourcin, A. J. and Moore, B. C. J. (1980). Lipreading connected discourse with fundamental frequency information. *British Society of Audiology Newsletter*, Summer Issue, August 1980, 42–43.

Sachs, M. B. and Young, E. D. (1979). Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate. *J. Acoust. Soc. Am.* **66**, 470–479.

Scheffers, M. T. M. (1984). Discrimination of fundamental frequency of synthesized vowel sounds in a noise background. *J. Acoust. Soc. Am.* **76**, 428–434.

Schindler, R. A. and Merzenich, M. M. (1985) (eds). *Cochlear Implants*, Raven Press, New York.

Schouten, J. F. (1968). The perception of timbre. *IPO Annual Progress Report* **3**, 32–34, Institute for Perception Research, Eindhoven.

Schroeder, M. R. (1959). New results concerning monaural phase sensitivity. *J. Acoust. Soc. Am.* **34**, 1579 (abstract).

Seneff, S. (1984). Pitch and spectral estimation of speech based on auditory synchrony model. Working Papers Vol. IV, Speech Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology.

Shankweiler, D., Strange, W. and Verbrugge, R. (1977). Speech and the problem of perceptual constancy. In *Perceiving, Acting, and Knowing: Toward an Ecological Psychology* (eds R. E. Shaw and J. Bransford), Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Shannon, R. V. (1983). Multichannel electrical stimulation of the auditory nerve in man. I. Basic psychophysics. *Hearing Res.* **11**, 157–189.

Shepard, R. N. (1972). Psychological representation of speech sounds. In *Human Communication: A Unified View* (eds E. E. David, Jr. and P. B. Denes), McGraw-Hill, New York.

Shinn, P. and Blumstein, S. E. (1984). On the role of the amplitude envelope for the perception of [b] and [w]. *J. Acoust. Soc. Am.* **75**, 1243–1252.

Sidwell, A. and Summerfield, Q. (1985). The effect of enhanced spectral contrast on the internal representation of vowel-shaped noise. *J. Acoust. Soc. Am.* **78**, 495–506.

Slawson, A. W. (1968). Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency. *J. Acoust. Soc. Am.* **43**, 87–101.

Soli, S. D. (1983). The role of spectral cues in discrimination of voice onset time differences. *J. Acoust. Soc. Am.* **73**, 2150–2165.

Srulovicz, P. and Goldstein, J. L. (1983). A central spectrum model: a synthesis of auditory-nerve timing and place cues in monaural communication of frequency spectrum. *J. Acoust. Soc. Am.* **73**, 1266–1276.

Stelmachowicz, P. G., Jesteadt, W., Gorga, M. P. and Mott, J. (1985). Speech perception ability and psychophysical tuning curves in hearing-impaired listeners. *J. Acoust. Soc. Am.* **77**, 620–627.

Stevens, K. N. (1952). Frequency discrimination for damped waves. *J. Acoust. Soc. Am.* **24**, 76–79.

Stevens, K. N. and Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In *Perspectives on the Study of Speech* (eds P. D. Eimas and J. L. Miller), Lawrence Erlbaum, Hillsdale, New Jersey.

Stevens, K. N. and Klatt, D. H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. *J. Acoust. Soc. Am.* **55**, 653–659.

Stock, D. and Rosen, S. (1986). Frequency discrimination and resolution at low frequencies in normal and hearing-impaired listeners: Preliminary results. *Speech, Hearing and Language: Work in Progress*, Phonetics and Linguistics, University College London (in press).

Summerfield, Q., Foster, J., Tyler, R. and Bailey, P. J. (1985). Influences of formant bandwidth and auditory frequency selectivity on identification of place of articulation in stop consonants. *Speech Commun.* **4**, 213–229.

Titze, I. R. and Talkin, D. T. (1979). A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation. *J. Acoust. Soc. Am.* **66**, 60–74.

Tong, Y. C., Millar, J. B., Clark, G. M., Martin, L. F., Busby, P. A. and Patrick, J. F. (1980). Psychophysical and speech perception studies on two multiple channel cochlear implant patients. *J. Laryngol. Otol.* **94**, 1241–1256.

Turek, S. Van de G., Dorman, M. F., Franks, J. R. and Summerfield, Q. (1980). Identification of synthetic /bdg/ by hearing-impaired listeners under monotic and dichotic formant presentation. *J. Acoust. Soc. Am.* **67**, 1031–1039.

Turner, C. W. and Van Tasell, D. J. (1984). Sensorineural hearing loss and the discrimination of vowel-like stimuli. *J. Acoust. Soc. Am.* **75**, 562–565.

Tyler, R. S., Wood, E. J. and Fernandes, M. (1982). Frequency resolution and hearing loss. *Br. J. Audiol.* **16**, 45–63.

Tyler, R. S., Wood, E. J. and Fernandes, M. (1983). Frequency resolution and discrimination of constant and dynamic tones in normal and hearing-impaired listeners. *J. Acoust. Soc. Am.* **74**, 1190–1199.

Tyler, R. S., Hall, J. W., Glasberg, B. R., Moore, B. C. J. and Patterson, R. D. (1984). Auditory filter asymmetry in the hearing impaired. *J. Acoust. Soc. Am.* **76**, 1363–1368.

Umeda, N. (1975). Vowel duration in American English. *J. Acoust. Soc. Am.* **58**, 434–445.

Velmans, M. (1973). Speech imitation in simulated deafness, using visual cues and 'recoded' auditory information. *Lang. Speech* **16**, 224–236.

Velmans, M. (1974). The design of speech recoding devices for the deaf. *Br. J. Audiol.* **8**, 1–15.

Velmans, M. and Marcuson, M. (1983). The acceptability of spectrum-preserving and spectrum-destroying transposition to severely hearing-impaired listeners. *Br. J. Audiol.* **17**, 17–26.

Voigt, H. F., Sachs, M. B. and Young, E. D. (1981). Effects of masking noise on the representation of vowel spectra in the auditory nerve. In *Neural Mechanisms of Masking* (eds J. Syka and L. Aitkin), Plenum Publishing, London.

Walden, B. E. (1984). Speech perception of the hearing-impaired. In *Hearing Disorders in Adults* (ed. J. Jerger), College-Hill Press, San Diego.

Wardrip-Fruin, C. (1982). On the status of temporal cues to phonetic categories: Preceding vowel duration as a cue to voicing in final stop consonants. *J. Acoust. Soc. Am.* **71**, 187–195.

Wardrip-Fruin, C. (1985). The effect of signal degradation on the status of cues to voicing in utterance-final stop consonants. *J. Acoust. Soc. Am.* **77**, 1907–1912.

Wells, J. C. (1982). *Accents of English* 3 vols, Cambridge University Press, Cambridge.

White, M. W. (1983). Formant frequency discrimination and recognition in subjects implanted with intracochlear stimulating electrodes. In *Cochlear Prostheses, an International Symposium* (eds C. W. Parkins and S. W. Anderson), *Ann. N.Y. Acad. Sci.* 405.

Wier, C. C., Jesteadt, W., and Green, D. M. (1977). Frequency discrimination as a function of frequency and sensation level. *J. Acoust. Soc. Am.* **61**, 178–184.

Wightman, F. L. (1982). Psychoacoustic correlates of hearing loss. In *New Perspectives on Noise-induced Hearing Loss* (eds R. P. Hamernik, D. Henderson and R. Salvi), Raven Press, New York.

Williams, L. (1977). The voicing contrast in Spanish. *J. Phonetics* **5**, 169–184.

Woolf, N. K., Ryan, A. F. and Bone, R. C. (1981). Neural phase-locking properties in the absence of cochlear outer hair cells. *Hearing Res.* **4**, 335–346.

Young, E. D. and Sachs, M. B. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. *J. Acoust. Soc. Am.* **66**, 1381–1403.

Zwicker, E. (1970). Masking and psychological excitation as consequences of the ear's frequency analysis. In *Frequency Analysis and Periodicity Detection* (eds R. Plomp and G. F. Smoorenburg), A. W. Sijthoff, Leiden.

Zwicker, E. and Schorn, K. (1982). Temporal resolution in hard-of-hearing patients. *Audiology* **21**, 474–492.

Zwicker, E. and Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am.* **68**, 1523–1525.