PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2008) **363**, 965–978
doi:10.1098/rstb.2007.2153
Published online 7 September 2007

Acoustic and auditory phonetics: the adaptive design of speech sound systems

Randy L. Diehl*

Department of Psychology and Center for Perceptual Systems, University of Texas at Austin, 1 University Station A8000, Austin, TX 78712, USA

Speech perception is remarkably robust. This paper examines how acoustic and auditory properties of vowels and consonants help to ensure intelligibility. First, the source–filter theory of speech production is briefly described, and the relationship between vocal-tract properties and formant patterns is demonstrated for some commonly occurring vowels. Next, two accounts of the structure of preferred sound inventories, quantal theory and dispersion theory, are described and some of their limitations are noted. Finally, it is suggested that certain aspects of quantal and dispersion theories can be unified in a principled way so as to achieve reasonable predictive accuracy.

Keywords: acoustic phonetics; auditory phonetics; speech sounds

1. INTRODUCTION

Speech sounds tend to be accurately perceived even in unfavourable listening conditions. Moore (2008) discusses several aspects of basic auditory processing that contribute to this perceptual robustness. The present paper considers how acoustic and auditory properties of commonly occurring speech sounds also help to ensure high levels of intelligibility. First, the source–filter theory of speech production is outlined, and the relationship between vocal-tract (VT) cavity size and shape and formant patterns is illustrated. Second, two theories intended to account for cross-language preferences in sound inventories, quantal theory and dispersion theory, are described and evaluated. Finally, it is suggested that a version of dispersion theory that incorporates certain aspects of quantal theory may have greater predictive success than either theory in its original form.

2. SOURCE–FILTER THEORY OF SPEECH PRODUCTION

The mapping between VT properties and acoustic signals has been investigated over many decades (Chiba & Kajiyama 1941; Stevens & House 1955, 1961; Fant 1960; Flanagan 1972; Stevens 1998) and, as documented in the last of these cited works, is now reasonably well understood for the major classes of speech sounds. At the core of this understanding lies the assumption that speech outputs can be analysed as the response of a set of VT filters to one or more sources of sound energy. A further assumption, that holds to a first approximation in most cases, is that the source and filter properties of the vocal tract are independent.

A source in the vocal tract is any modulation of the airflow that creates audible energy. Such sound-producing modulations occur in the vicinity of constrictions either at the glottis (i.e. the space between the vocal folds of the larynx) or in the supralaryngeal regions of the vocal tract. Several types of source may be distinguished. One is (quasi-) periodic and consists of cycles of varying airflow attributable to vocal-fold vibration or voicing. Sounds produced with a voiced source have a fundamental frequency (F0) equal to the repetition rate of vocal-fold vibration. They include vowels (e.g. /a/ and /u/), nasal consonants (e.g. /m/), liquids (e.g. /r/ and /l/) and glides (e.g. /w/). Other sources are aperiodic and include (i) turbulence noise generated as air flows rapidly through an open, non-vibrating glottis (referred to as 'aspiration'), (ii) turbulence noise generated as air flows rapidly through a narrow supralaryngeal constriction (referred to as 'frication'),[1] and (iii) a brief pulse of excitation caused by a rapid change in oral air pressure (referred to as a 'transient source'). Examples of the use of these aperiodic sources are, respectively, the aspirated /h/, the fricatives /f/ and /s/ and the stop consonants /p/ and /t/ (both of which, in stressed-syllable-initial position, tend to be associated with a rapid reduction in oral air pressure at the moment of VT opening). Some speech sounds have multiple sources operating simultaneously or in succession. For example, the fricative /z/ is produced with a voiced source and a simultaneous turbulence noise (i.e. frication) source, while the stop consonant /t/ may be produced with, in quick succession, a transient source, a frication source and an aspiration source, as the mouth opens (Fant 1973).

All of these sources—both periodic and aperiodic—are well suited for evoking responses from the VT filters. Under normal conditions, each source has an energy level sufficient to generate highly audible speech sounds. Moreover, each source has an amplitude spectrum that is fairly broadband, ensuring that even VT filters in the higher-frequency range (1–5 kHz) will tend to be excited.

*diehl@psy.utexas.edu

One contribution of 13 to a Theme Issue 'The perception of speech: from sound to meaning'.

This journal is © 2007 The Royal Society

How, then, does the vocal tract act to filter sound energy generated by the sources? Any fully or partially enclosed volume of air has certain natural frequencies of vibration, or resonance frequencies, that are determined mainly by the size and shape of the volume, and by the extent and character of the enclosing surfaces. When the volume of air is exposed to a broadband energy source, it will respond strongly to source frequencies at or near its resonance frequencies and weakly to other source frequencies. This relative response as a function of frequency defines the filter, or transfer, function of the vocal tract in a given configuration.[2]

Figure 1 shows the source–filter theory for four different vowel sounds. Figure 1*a*(i) displays an idealized spectrum of the glottal airflow waveform corresponding to a voiced source. The value of F0 is 100 Hz, and the slope of this spectrum is −12 dB per octave, values typical of an adult male voice. Since efficiency of sound transmission from the mouth (known as the 'radiation characteristic') increases at frequencies above 300–500 Hz at a rate of 6 dB per octave, the effective glottal spectrum slope with respect to the listener is −6 dB per octave (see dotted curve). Figure 1*b* shows filter functions for the vowels /ə/, or schwa, (as in the first syllable of 'about'), /u/ (as in 'boot'), /i/ (as in 'beet') and /a/ (as in American English 'hot'), with each function including three resonance peaks within the 0–3 kHz range. Figure 1*c* represents the acoustic output spectra of the four vowels. On the assumption of source–filter independence, the spectrum of the output sound is considered to be the product of three terms: the source spectrum; the VT filter function; and the radiation characteristic (Stevens & House 1961).

An important consequence of the near independence of VT sources and filters is that the speech signal can transmit linguistic information at higher rates than would otherwise be possible. For example, in most languages, F0 variations are used to convey lexical, grammatical and paralinguistic (e.g. attitudinal or emotional) information in parallel with that provided by the sequencing of vowels and consonants. Moreover, consonant sounds in most languages are distinguished on the basis of both source and filter properties of the vocal tract (Maddieson 1984). A significant degree of independence also characterizes the relationship between different VT sources (e.g. voicing and frication) and between different VT filter properties (e.g. place of articulation and nasality, see later), further increasing the information content of speech. This principle of independence is one factor underlying the application of the descriptive framework of distinctive features in the study of the world's languages (Jakobson *et al.* 1963; Chomsky & Halle 1968; Diehl & Lindblom 2004).

3. VOCAL-TRACT CAVITY PROPERTIES AND FORMANT FREQUENCIES

A key concept in acoustic phonetics is the 'formant'. It refers to the acoustic realization of an underlying resonance peak in the VT filter function and is illustrated by the envelope peaks in the output spectra of each of the vowels represented in figure 1. A formant is characterized by a centre frequency, a relative amplitude and a bandwidth. For the acoustic description of vowel sounds, the most important parameters are the centre frequencies of the lowest three or four formants, referred to as the 'formant pattern' collectively. Perceived vowel identity (e.g. whether a vowel token is heard as an instance of /i/ or /u/) is strongly influenced by the formant pattern but only modestly affected (across a sizable range of values) by the relative amplitudes or bandwidths of the formants (Klatt 1982). Given any formant pattern and a glottal source spectrum, the acoustic theory of vowel production (Fant, 1960, 1973; Stevens & House 1961) makes reasonably accurate predictions about formant bandwidths and relative amplitudes and, hence, the overall shape of the spectral envelope.

To understand the relationship between the size and shape of the vocal tract and the formant pattern, consider first the vowel /ə/ (the top-most vowel represented in figure 1). During production of this vowel, the cross-sectional area of the vocal tract is approximately uniform from the region just above the glottis all the way to the lips. This uniform tube is open at the lips and effectively closed at the glottal end because the average size of the glottis during vocal-fold vibration is very small relative to the cross-sectional area of the supralaryngeal vocal tract. When a pressure wave is generated by airflow through the vibrating vocal folds, it travels to the lip opening where it is almost completely reflected owing to the very small impedance outside the mouth. There is a boundary condition of essentially zero acoustic pressure at the lips. For each frequency component in the source, the corresponding forward-going wave from the glottis and the reflected wave from the lips combine to form a standing wave with a wavelength inversely related to frequency. The amplitude of the standing wave varies sinusoidally along the length of the vocal tract, with nodes (i.e. zero points corresponding to steady atmospheric pressure) located at the lips and at every half-wavelength back to the glottis and antinodes (points of maximum positive and negative deviations from atmospheric pressure) located at odd multiples of the quarter-wavelength distance from the lips. Each standing pressure wave has a corresponding standing volume–velocity (airflow) wave with nodes and antinodes located, respectively, at the antinodes and nodes of the standing pressure wave. Resonance occurs at just those frequencies for which a pressure antinode (a volume–velocity node) occurs at the glottal end of the vocal tract.[3]

Figure 2 shows the standing pressure waves for the three lowest resonance frequencies (500, 1500 and 2500 Hz) of /ə/, given a VT length, l, of 17.5 cm, a typical adult male value. Notice that in each case the boundary conditions described in the previous paragraph are met. The lowest frequency resonance, corresponding to the first formant, is represented at the bottom. It may be observed that the standing pressure wave extending from the glottis to the lips amounts to one-quarter of a sinusoidal cycle; thus, the wavelength, λ, equals $4l$ or 70 cm. Accordingly, tubes closed at one end and open at the other end are called 'quarter-wave resonators'. The corresponding resonance or formant frequency, F1, is calculated from

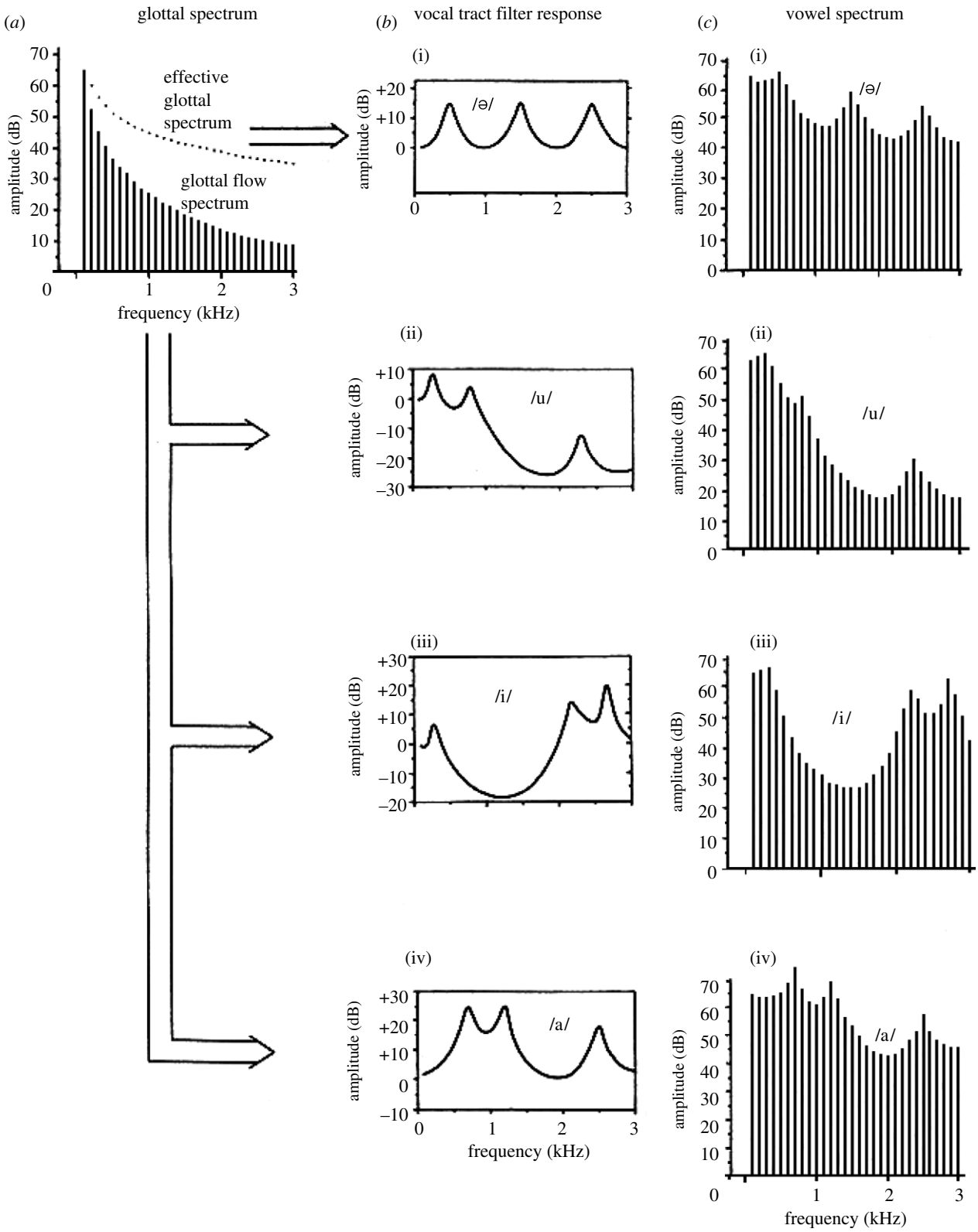

Figure 1. Source–filter theory of speech production illustrated for the vowels /ə/, /u/, /i/ and /a/. (*a*) An idealized spectrum of the glottal airflow waveform, with a slope of −12 dB per octave, is displayed. The effective glottal spectrum slope (dotted curve) is −6 dB per octave owing to more efficient sound transmission from the mouth at higher frequencies. (*b*) Filter functions for the four vowels. (*c*) Product of the glottal source spectrum and the filter functions yields the acoustic output spectra. (Adapted with permission from Pickett (1999), Allyn & Bacon; adapted from Fant (1960) and Stevens & House (1961).)

the formula $f=c/\lambda$ (where c is equal to the speed of sound, approx. 35 000 cm s^{-1}), yielding a value of 500 Hz. Analogous calculations for the second and third formant frequencies (F2 and F3) give values of 1500 and 2500 Hz. These frequencies are consistent with the resonance and formant peaks shown for /ə/ in figure 1.

Models of VT configurations for vowels other than /ə/ require either tubes of non-uniform cross-sectional area or else a series of two or more uniform tubes with different cross-sectional areas. In the case of the vowel /a/, the jaw and tongue body are lowered, creating a large oral (front) cavity, and the tongue is also

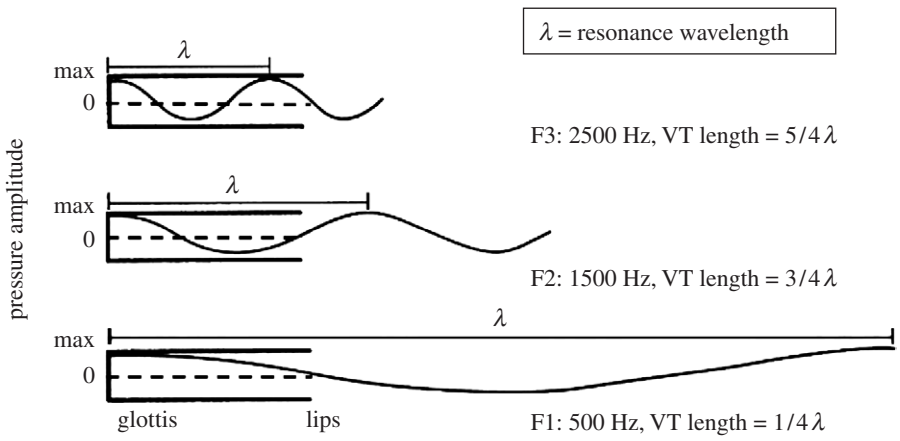

Figure 2. The standing pressure waves for the three lowest resonance frequencies (500, 1500, 2500 Hz) of the vowel /ə/, produced with a vocal-tract (VT) length of 17.5 cm. Each standing wave satisfies the boundary conditions that an antinode exists at the closed (glottal) end of the vocal tract and a node exists at the open (lip) end. F1, F2 and F3 refer to the first three formants, corresponding to the first three resonances of the vocal tract. (Adapted with permission from Johnson (1997), Blackwell Publishers.)

retracted, creating a narrow pharyngeal (back) cavity. This configuration can be modelled as a series of two quarter-wave resonators, that is, a series of two uniform tubes each effectively closed at the input end and open at the output end. The back cavity is treated as open at its output end, while the front cavity is treated as closed at its input end owing to the relatively large difference in cross-sectional area between the two cavities. The size of this difference also implies that the two tubes can be considered acoustically independent, at least to a first approximation. This means that the filter function for the entire VT configuration can be estimated by combining the separate resonance frequencies of the front and back cavities. Each of the two quarter-wave resonators used in the production of /a/ is, of course, shorter than the single one used to produce /ə/, and thus their lowest resonance frequencies are higher in value. In addition, the two resonators are comparable in length, yielding F1 and F2 values that are relatively close together. These acoustic properties of /a/ are shown in figure 1.

In the case of the vowel /i/, the jaw and tongue body are raised, creating a narrow front cavity, and the tongue body is moved forward, enlarging the back cavity. This configuration can be modelled as a series of two uniform tubes with very different cross-sectional areas.[4] The wide back cavity is effectively closed at both the glottal end and the forward end that communicates with the narrow front cavity, while the front cavity is open at both ends. For a uniform tube closed at both ends, resonance occurs only at frequencies for which there are antinodes in the standing pressure wave at the closed ends of the tube and a node at the very middle of the tube. If a tube is open at both ends, resonance occurs only at frequencies for which there is a node at each end and an antinode at the middle. In both cases, the lowest frequency standing wave extending across the length of the tube is one-half of a sinusoidal wavelength, and the tubes are thus referred to as 'half-wave resonators'. Other things being equal, half-wave resonators have a lowest natural frequency that is double that of quarter-wave resonators. In figure 1, the relatively high F2 and F3 of /i/ correspond to the lowest frequency resonances of the front and back tubes. The low F1 of this vowel is attributable to a 'Helmholtz resonator' comprising both the wide back and the narrow front cavities. The natural frequency of a Helmholtz resonator increases with the square root of the cross-sectional area of the front cavity, and decreases with the square root of the length of the front cavity and the volume of the back cavity. Given the dimensions of VT cavities, such a resonator has a low natural frequency.

For the vowel /u/, the tongue body is raised and retracted, producing a wide front cavity and a narrow constriction between the front and back cavities, the tongue root (the lower back portion of the tongue that forms the front wall of the pharynx) is moved forward, creating a wide back cavity, and the lips are rounded, creating a narrow opening. This configuration can be modelled both as a series of four uniform tubes (wide–narrow–wide–narrow), all of which are half-wave resonators, and as a series of two Helmholtz resonators. As shown in figure 1, /u/ has a low F1 and F2 which are produced by the two Helmholtz resonators, and a high F3 which is produced by the longest of the half-wave resonators.

More realistic non-uniform tube models based on the accurate measurements of VT dimensions are, of course, possible. Nevertheless, simplified models consisting of uniform tubes and Helmholtz resonators suffice to illustrate some of the main principles underlying the relationship between VT properties and formant patterns.

4. ADAPTIVE DESIGN OF SPEECH SOUND INVENTORIES

(a) *The restricted character of speech sound systems*

Among the vowels and consonants that have been observed in the world's languages, some occur commonly, whereas most are relatively rare (Crothers 1978; Maddieson 1984). What factors might explain such cross-language preferences for certain sounds over others? One possible factor, long discussed by linguists (Passy 1890; Jakobson 1941; Martinet 1955), is the requirement that speech sounds be audible and

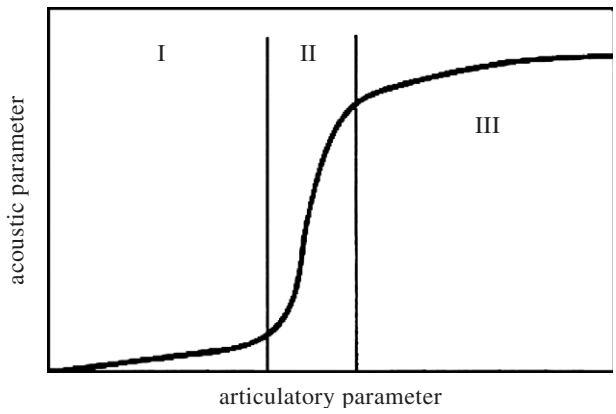

Figure 3. A schematic of a quantal nonlinearity in the mapping between an articulatory parameter of the vocal tract and the acoustic output. Regions I and III are acoustically quite stable with respect to perturbations in the articulatory parameter, whereas region II is acoustically unstable. Speech sound categories are assumed to be located in regions I and III. (Adapted with permission from Stevens (1989), Academic Press.)

distinctive (i.e. not confusable with other speech sounds). A second possible factor is a general tendency towards efficiency in human behaviour (Zipf 1949) such that goals—in this case, successful speech communication—are achieved with minimum effort. These two factors will be referred to, respectively, as 'listener-oriented' and 'talker-oriented' constraints on sound selection.

During the last several decades, two theories, quantal theory (Stevens 1972, 1989, 1998) and dispersion theory (Liljencrants & Lindblom 1972; Lindblom 1986; Diehl *et al.* 2003; Diehl & Lindblom 2004), have been developed to account for cross-language preferences in the structure of sound inventories. The two theories are broadly similar in emphasizing both listener- and talker-oriented selection factors, but they differ in their characterization of these factors.

(b) *Quantal theory: vowels*

Quantal theory is based on the observation that certain nonlinearities exist in the mappings between articulatory (i.e. VT) configurations of talkers and acoustic outputs and also between the speech signals and the auditory responses of listeners. Such a nonlinearity in the articulatory-to-acoustic mapping is represented in figure 3. In regions I and III, perturbations in the articulatory parameter result in small changes in the acoustic output, whereas in region II, comparably sized perturbations yield large acoustic changes. Given these alternating regions of acoustic stability and instability, an adaptive strategy for a language community is to select sound categories that occupy the stable regions and that are separated by the unstable region. Locating sound categories in stable regions allows talkers to achieve an acceptable acoustic output with less articulatory precision than would otherwise be necessary, thus helping to satisfy the talker-oriented goal of minimum effort. In addition, separating two sound categories by a region of acoustic instability ensures that they are acoustically very different, and thus helps to satisfy the listener-oriented requirement of sufficient

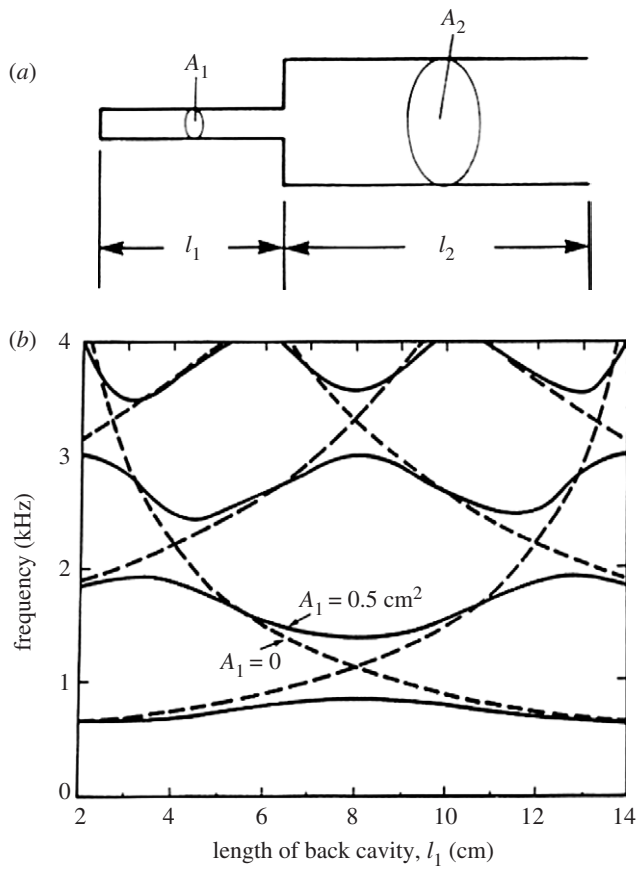

Figure 4. (*a*) A two-tube model of the vocal tract, with each tube effectively closed at the input end and open at the output end. The lengths of the left (back) and right (front) tubes are l_1 and l_2, and the cross-sectional areas are A_1 and A_2, respectively. (*b*) A nomogram representing the first four resonance frequencies for the two-tube model in (*a*) as the length l_1 of the back cavity is varied, with overall length l_1+l_2 held constant at 16 cm and $A_2=3$ cm^2. The dashed curves represent the case where $A_1 \gg A_2$; the solid curves represent the case where $A_1=0.5$ cm^2. (Adapted with permission from Stevens (1989), Academic Press.)

auditory distinctiveness. Another advantage for listeners is that vowels produced in acoustically stable regions should be relatively invariant. According to quantal theory, this convergence of talker- and listener-oriented selection criteria leads to a preference for certain 'quantal' vowels and consonants.

Consider, for example, the VT model in figure 4*a* consisting of a series of two quarter-wave resonators with lengths l_1 (back cavity) and l_2 (front cavity) and cross-sectional areas A_1 and A_2. (See the earlier discussion of the two-tube model for the vowel /a/.) Figure 4*b* is a nomogram representing the effects on the first four resonance frequencies of varying l_1 while the total length of the model, l_1+l_2, is held constant at 16 cm and $A_2=3$ cm^2. If the ratio of A_1 to A_2 is very small, the resonance frequencies of one tube are roughly independent of those of the other. The idealized case of complete independence is represented by the dashed curves in the nomogram. The resonance frequency curves with an upward trend as a function of l_1 are associated with the front cavity; those with a downward trend are associated with the back cavity. Note that the correspondence between cavities and

numbered formants in the acoustic output changes at the crossover points in the curves.

The solid curves in figure 4*b* represent resonance frequencies of the two-tube model when A_1 is increased to 0.5 cm^2, a value large enough to yield the non-turbulent airflow characteristic of vowels. The larger ratio of A_1 to A_2 in this case causes a modest degree of acoustic coupling between the two cavities such that the resonance frequencies of each cavity are influenced by the other. A main result is that the curves no longer intersect but rather approach each other and then diverge. In the region of convergence, perturbations in l_1 have small effects on formant frequencies, whereas in adjacent regions such perturbations have larger effects. It turns out that at an l_1 value of approximately 8 cm, with maximum convergence between F1 and F2 and also between F3 and F4, the two-tube model corresponds rather closely to the VT configuration for the vowel /a/.

Stevens (1972, 1989) used this example to illustrate the notion of a quantal vowel and presented similar arguments with respect to the vowels /i/ and /u/. Recall that, as in the case of /a/, /i/ is produced with a large difference in cross-sectional area between the front and back cavities (but in the opposite direction), whereas /u/ is produced with a narrow constriction separating the back and front cavities. For all three vowels, therefore, the front and back cavities are only weakly coupled acoustically; however, the coupling is sufficient to yield acoustically stable quantal regions that alternate with acoustically unstable regions, where stability is defined with respect to variation in cavity length parameters. As may be inferred from these examples, weak yet non-negligible coupling between different VT resonators is a necessary (though plainly not a sufficient) condition for quantal vowel status. Zero coupling would produce no regions of acoustic stability (see the dashed curves in figure 4*b*), while a high degree of coupling would flatten the peaks and troughs of the resonance curves, creating large stable regions but without the adjacent regions of acoustic instability that confer auditory distinctiveness (Diehl 1989).

Stevens (1972, 1989) emphasized another property of quantal vowels that should be advantageous for listeners. When two resonances are in close proximity, they reinforce each other creating a relatively intense spectral region. This may be confirmed by examining the filter functions and output spectra for the non-schwa vowels in figure 1. For /u/ and /a/ there are prominences in the low- and mid-frequency regions, respectively, owing to the convergence of F1 and F2, whereas for /i/ there is a similar prominence in the high-frequency region due to the convergence of F2 and F3.

To summarize, quantal vowels appear to satisfy listener-oriented selection criteria in at least three ways: (i) being produced in acoustically stable regions of phonetic space, they are relatively invariant, (ii) being separated from nearby vowels by regions of acoustic instability, their formant patterns are auditorily distinctive, and (iii) being characterized by relatively intense spectral prominences, they are likely to be resistant to masking by background noise (Darwin 2008; Moore 2008).

A common misinterpretation of quantal theory is that it predicts greater stability for quantal vowels such as /i/ or /u/ than for non-quantal vowels like /ʌ/ (as in American English 'cup'), which is produced without a major VT constriction (Ladefoged *et al.* 1977; Pisoni 1980; Syrdal & Gopal 1986). Several results inconsistent with this putative prediction have been reported. For example, Pisoni (1980) found that when talkers repeatedly mimicked synthetic vowel sounds, within-talker variances in F2 were smaller for /ʌ/ than for /i/ or /u/.

However, as Diehl (1989) noted, quantal theory does not, in fact, predict that a quantal vowel is more stable than *any* non-quantal vowel. Rather, as shown in figure 4*b*, it predicts that a quantal vowel will be more stable than any non-quantal vowel that is produced with different cavity length parameters but with the same cross-sectional area parameters. A vowel like /ʌ/ is actually predicted to be among the most stable of vowels by quantal theory (with respect to perturbations in l_1) because the back and front cavities differ little in cross-sectional area and are, therefore, highly coupled acoustically. Recall that greater acoustic coupling results in a flattening of the peaks and troughs of the resonance frequency curves in the nomogram. (When A_1 and A_2 are equal, creating a uniform tube corresponding to schwa, variation in 'l_1' obviously does not alter the configuration at all, and the frequency curves become perfectly horizontal.) Accordingly, the vowel /ʌ/ lacks quantal status not because it occupies an unstable region of phonetic space, but because it is not bounded by regions of acoustic instability that confer auditory distinctiveness vis-à-vis other nearby vowels. In other words, /ʌ/ satisfies the talker-oriented, but not the listener-oriented, selection criteria of quantal theory.

Although quantal theory is not falsified by the evidence that non-quantal vowels like /ʌ/ are more stable than quantal vowels such as /i/ or /u/, the theory nevertheless faces a major difficulty regarding the claim that the quantal vowels are relatively stable. Recall that acoustic stability is defined in quantal theory with respect to variation in some cavity length parameter such as l_1 in figure 4*a*. However, it is reasonable to ask how stable quantal vowels are with respect to variations in other VT parameters such as A_1. In figure 4*b*, it may be seen that the largest effects on resonance frequencies of varying the A_1 parameter occur in just those regions that are most stable with respect to perturbations in l_1. Thus, quantal vowels are actually the *least* stable vowels with respect to changes in cross-sectional area parameters. This would perhaps not be a serious problem for quantal theory if the acoustic effects of varying cavity width were relatively small. However, as noted by Diehl (1989), perturbations in the width of a VT cavity tend to yield changes in resonance frequencies that are at least equal to—and often greater than—those caused by comparable perturbations in cavity length. The relative stability of quantal vowels is, therefore, questionable.

This argument in no way undermines quantal theory's claims that quantal vowels are favoured by the listener-oriented criteria of auditory distinctiveness and audibility in noise. It is possible that these criteria alone may be sufficient to generate accurate predictions about cross-language preferences in the structure of

vowel inventories. In the UCLA Phonological Segment Inventory Database (UPSID, Maddieson 1984) of 317 diverse languages, the most commonly occurring vowels are /i/, /a/ and /u/, which appear in 92, 88 and 84% of the languages, respectively. As was discussed earlier, each of these vowels clearly meets the listener-oriented criteria for quantal status and, accordingly, their high frequency of occurrence is consistent with quantal theory.

However, /i/, /a/ and /u/ are not the only quantal vowels. When a tongue configuration appropriate for /i/ is combined with lip rounding, the resulting vowel is /y/ (as in the first syllable of the German word 'über'). Relative to /i/, both F2 and F3 are shifted downward for this vowel, but otherwise the nomogram is very similar (Stevens 1989). Both vowels have closely spaced values of F2 and F3 and each is produced in a stable region of the l_1 dimension that is bounded by unstable regions. (For /y/ the stable region occurs at a somewhat higher value of l_1.) If /i/ is a quantal vowel, so too is /y/. Presumably, then, quantal theory would predict a frequency of occurrence for /y/ that is comparable to that for /i/. However, /y/ occurs in only approximately 8% of the languages of the UPSID sample (Maddieson 1984) and is virtually absent in languages with small vowel inventories. The large discrepancy in frequency of occurrence between /i/ and /y/ is difficult to explain within the framework of quantal theory.

After /i/, /a/ and /u/, the most commonly occurring vowels in the UPSID sample (Maddieson 1984) are the mid-front vowels /e/ (as in Spanish 'tres') or /ɛ/ (as in English 'bet') and the mid-back vowels /o/ (as in Spanish 'dos') or /ɔ/ (as in American English 'bought'). Given the degree of VT constriction during their production and the proximity of their F1 and F2 values, /o/ and /ɔ/ appear to be quantal vowels, and their high frequency of occurrence is thus predicted by quantal theory. However, the same is not true for vowels /e/ and /ɛ/. During the production of these vowels, the vocal tract is relatively unconstricted (Fant 1960; Perkell 1979) and, as in the case of /ʌ/, such vowels cannot be considered quantal because they lack surrounding regions of acoustic instability that yield auditory distinctiveness.

Quantal theory thus has only mixed success as an account of preferred vowel inventories. Although it correctly predicts the frequent occurrence across languages of /i/, /a/, /u/ and /o/ (or /ɔ/), it fails to predict the high frequency of /e/ (or /ɛ/) and the low frequency of /y/.

(c) *Quantal theory: consonants*

Most of Stevens's work on quantal theory has focused on vowel sounds; however, the quantal notion is intended to apply also to certain consonant sounds and to the distinction between vowels and consonants. Thus, for example, there are quantal contrasts between stop consonants (with the airflow completely blocked across the region of articulatory closure), fricative consonants (with the articulatory constriction sufficient to produce turbulence noise) and vowel-like sounds (i.e. liquids, glides and true vowels, all of which have sufficient articulatory opening to produce mainly laminar, or non-turbulent airflow; Stevens 1972). The quantal character of these contrasts is reflected in the nonlinear mapping between articulatory settings (e.g. cross-sectional area of the constriction) and the acoustic signal.

Contrasts based on varying degrees of VT constriction, such as those described in the previous paragraph, are often referred to as distinctions in *manner of articulation*. Other important consonant distinctions include: *oral* versus *nasal* (e.g. /b/ versus /m/, or /d/ versus /n/), reflecting whether the soft palate, or velum, is raised or lowered, the latter case resulting in acoustic coupling of the nasal and oral cavities; *place of articulation* (e.g. /b/ versus /d/ versus /g/, as in 'go'), corresponding to the location in the vocal tract where the most prominent articulatory closure or constriction occurs; and *voiced* versus *voiceless* (e.g. /b/ versus /p/, or /d/ versus /t/), indicating whether or not vocal-fold vibration occurs in the temporal vicinity of the consonant constriction and/or release.

The oral/nasal consonant distinction is quantal in a way analogous to the distinction between stop consonants and continuant consonants, such as fricatives and glides. In both cases, there is either complete occlusion of an airway or some degree of opening, and the difference between these two states can be modelled as a region of acoustic instability. Moreover, in both cases, the occluded state is acoustically stable with respect to a range of muscular forces, whereas the non-occluded state is relatively stable with respect to a range of constriction sizes. The occurrence of oral/nasal consonant contrasts in approximately 97% of the languages in the UPSID sample (Maddieson 1984) is, therefore, consistent with quantal theory.

Just as there are quantal regions for vowels along the back-cavity length (l_1) dimension, Stevens (1989) noted that there are several quantal regions for consonants along the place-of-articulation dimension. (As in the case of vowels, these regions correspond to intersections between resonance frequency curves such that two formants are in close proximity and are relatively stable in frequency with respect to perturbations in place of articulation.) These quantal regions occur at the velar place of articulation (i.e. the oral occlusion is between the tongue body and velum), where F2 and F3 are close together, and at the retroflex place of articulation (i.e. the tongue tip is raised and retracted to occlude the vocal tract at the hard palate), where F3 and F4 are close together. Velar stop consonants (/g/ or /k/) occur in more than 99% of languages in the UPSID sample (Maddieson 1984), which counts as a successful prediction of quantal theory. However, retroflex stops (e.g. /ɖ/ in Hindi) occur in only approximately 12% of languages in the UPSID sample. It is unclear how quantal theory can account for the differing frequencies of velar and retroflex stops without appealing to principles outside the theory.

Even more problematic for quantal theory is the high cross-language frequency (more than 99% of the languages in the UPSID sample, Maddieson 1984) of stops having a labial place of articulation (/b/ or /p/) and those having a dental/alveolar place of articulation (/d/ or /t/). ('Labial' refers to occlusion at the lips; 'dental' refers to occlusion between the tongue tip and the rear surfaces of the upper teeth; and 'alveolar' refers to occlusion between the tongue tip or blade and the

upper gum or alveolar ridge.) As noted by Diehl (1989), neither labials nor dentals/alveolars satisfy Stevens's criteria for quantal status. This is because the front-cavity resonances for these place values are too high in frequency to intersect with the back-cavity resonances within a perceptually significant frequency range, thus removing the possibility of acoustically stable regions with formants close together. (See Diehl (1989) for similar arguments with respect to fricative consonants.)

In all of the cases of possible quantal effects discussed so far, the putative nonlinearities occur in the mapping between articulatory configurations and acoustic outputs. Recall that Stevens (1989) noted that nonlinear relations between acoustic signals and auditory responses might also yield preferences for certain sound categories or sound category contrasts. There is evidence, for example, that an auditory nonlinearity (not specifically cited by Stevens) may help to account for the widespread use of consonant voicing contrasts among languages (61% of the UPSID sample, Maddieson 1984).

In a cross-language study of syllable-initial stops, Lisker & Abramson (1964) identified an important phonetic correlate of voicing contrasts, namely, *voice onset time* (VOT), the interval between the release of the articulators (e.g. opening of the lips) and the start of vocal-fold vibration. Across languages, stops in initial position tend to occur within three VOT sub-ranges: long negative VOTs (voicing onset precedes the articulatory release by more than 45 ms); short positive VOTs (voicing onset follows the release by no more than 20 ms); and long positive VOTs (voicing onset follows the release by more than 35 ms). From these three VOT sub-ranges, languages typically select two adjacent ones to implement their voicing contrasts. For example, Spanish and Dutch use long negative VOT values for their voiced category and short positive VOT values for their voiceless category, whereas English and Cantonese use the short positive VOT sub-range for their voiced category and the long positive VOT sub-range for their voiceless category. Thai is a rare example of a language that exploits all three VOT sub-ranges to implement a three-way voicing contrast.

Initial stops with negative VOT values have a low-frequency 'voice bar' during the closure interval that is followed, starting at the moment of articulatory release, by a broadband of mainly periodic energy concentrated in the formant regions. For stops with positive VOT values, there is no voice bar during the closure, and the VOT interval is characterized by a strongly attenuated first formant and by higher formants that are excited aperiodically.

Lisker & Abramson (1970; Abramson & Lisker 1970) examined perception of synthetic VOT syllables by native-speaking listeners of English, Spanish and Thai. They found that for each of the three language groups VOT perception is 'categorical' in the sense that listeners show (i) sharp identification boundaries between the categories relevant for their language, (ii) relatively good discrimination of stimulus pairs that straddle category boundaries and (iii) relatively poor discrimination of stimulus pairs drawn from the same voicing category. (This pattern of perceptual performance has been demonstrated for a variety of consonant contrasts. For reviews, see Repp (1984) and Diehl *et al.* (2004).)

Considered in isolation, categorical perception of speech sounds by adult human listeners does not provide convincing evidence for the existence of quantal effects based on auditory nonlinearities. Enhanced discriminability at category boundaries might simply reflect language-specific experience in categorizing speech sounds. Several lines of evidence, however, support the conclusion that discrimination peaks near voicing category boundaries are at least part of a general auditory character. First, infants from a Spanish-speaking environment showed enhanced discrimination of VOT differences that straddle either the Spanish or the English voicing boundaries (Lasky *et al.* 1975), and a similar pattern of results was found for infants from an English-speaking environment (Aslin *et al.* 1981). Second, adult listeners' discrimination functions for non-speech analogues of VOT stimuli exhibit similar regions of peak performance (Miller *et al.* 1976; Pisoni 1977). Third, non-human animals (chinchillas) that were first trained to respond differently to two endpoint stimuli from a synthetic alveolar VOT series (/da/, 0 ms VOT; and /ta/, 80 ms VOT), and then tested on the full series, showed identification functions very similar to those of English-speaking humans (Kuhl & Miller 1978).

A neural correlate of heightened discriminability near the English voicing category boundary was reported by Sinex *et al.* (1991). They recorded auditory nerve responses in chinchilla to stimuli from a VOT series (/da/–/ta/). For stimuli that were well within either the voiced or the voiceless category, there was high-response variability across neurons with different best frequencies. However, for the 30 ms VOT and 40 ms VOT stimuli, located near, but on opposite sides of, the English /d/–/t/ boundary, the response to the onset of voicing was highly synchronized across the same sample of neurons. This pattern of neural responses is shown in figure 5.

Consistent with quantal theory, the above findings suggest that in order to achieve enhanced distinctiveness, languages may exploit certain auditory nonlinearities in selecting sound categories.[5]

(d) *Dispersion theory: vowels*

As remarked earlier, the idea that speech sound inventories are structured to maintain perceptual distinctiveness has a long history in linguistics. However, the first investigators to express this idea quantitatively were Stevens (1972), in his first paper on quantal theory, and Liljencrants & Lindblom (1972), who took a quite different approach to the problem. Whereas in quantal theory distinctiveness characterizes the relationship between sound categories in localized regions of phonetic space (*viz.* where the categories are separated by acoustically unstable zones), Liljencrants & Lindblom viewed distinctiveness as a global property of an entire inventory of sound categories. A vowel or consonant inventory was said to be maximally distinctive if the sounds were maximally dispersed (i.e. separated from each other) in the available phonetic space.[6]

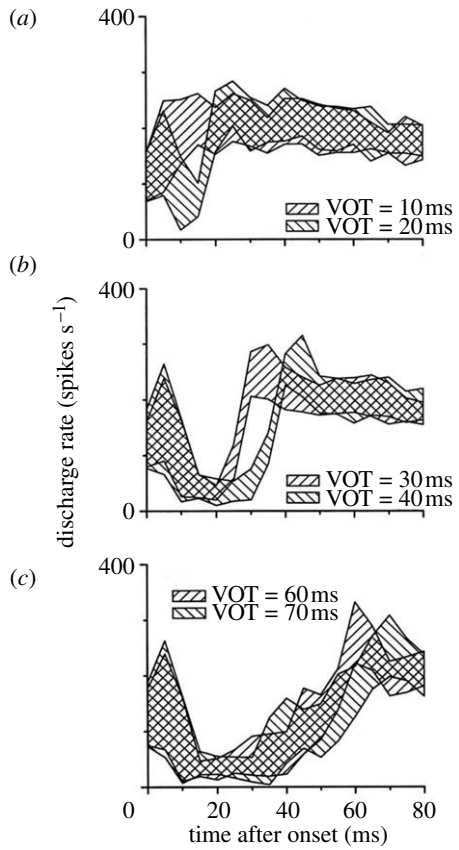

Figure 5. Auditory nerve responses in chinchilla to pairs of alveolar VOT stimuli in which the VOT difference was 10 ms. Each cross-hatched area encloses the mean $\pm$1 s.d. of the average discharge rates of neurons. (Adapted with permission of the first author from Sinex *et al.* (1991).)

Dispersion theory was primarily applied to vowel sounds, and predicted vowel inventories will be the main focus of discussion. Liljencrants & Lindblom (1972) began by specifying the available phonetic space from which particular vowel sounds might be selected. This space comprises the set of possible acoustic outputs of a computational model of the vocal tract (Lindblom & Sundberg 1971) that reflects natural articulatory degrees of freedom for the jaw, lips, tongue and larynx. The model outputs were restricted to simulations of non-nasal vowel-like sounds (i.e. sounds produced without coupling between the oral and nasal cavities and without a narrow constriction in the pharyngeal or oral portions of the vocal tract). The outputs were stationary in frequency and were represented as points in a Mel-scaled $F1 \times F2'$ space where $F2'$ corresponds to an effective F2 value corrected for the influence of F3. The $F1 \times F2'$ space was densely sampled at equal-Mel intervals to create a large set of candidate vowel sounds. (The Mel scale is a measure of subjective frequency similar to the Bark scale,[7] and also related to the ERB_N-number scale described by Moore (2008).)

To simulate preferred vowel inventories, Liljencrants & Lindblom (1972) next applied the following selection criterion: for any given size of vowel inventory, choose those candidate vowels that maximize pairwise Euclidean distances within the $F1 \times F2'$ space. The results of these simulations are shown in figure 6 for inventory sizes between 3 and 12. Solid curves represent the range of possible outputs from the articulatory model of Lindblom & Sundberg (1971), and filled circles correspond to the vowels selected according to the maximum distance criterion. (Note that the formant frequency axes are now scaled in kilohertz, though distances were calculated in Mel units.) Figure 6*d* pools all vowels selected across the 10 inventory sizes.

It is evident that a large majority of selected vowels are located on the periphery of the vowel space; central vowels appear only in the larger inventories. This pattern is consistent with the vowel inventory structure of most natural languages (Crothers 1978; Maddieson 1984). To evaluate in more detail how well dispersion theory predicts favoured vowel inventories, consider the simulations for the 3-, 5- and 7-vowel systems. (The vowels represented by the filled circles in these three cases may be identified by referring to figure 7.) For the 3-vowel inventory in figure 6, the predicted system includes /i/, /u/ and a slightly fronted version of /a/. Recall that these vowel categories are the most frequently occurring among the world's languages, and they all appear in the most common 3-vowel inventories (Crothers 1978; Maddieson 1984). Given their locations at the extreme points of the vowel space, it is not at all surprising that these vowels would be selected by a criterion of maximum dispersion.

For the 5-vowel inventory, the predicted system in figure 6 again includes /i/ and /u/ and an even more fronted version of /a/ (approaching /æ/, as in American English 'bat'). In addition, this system includes a mid-front vowel between /e/ and /ɛ/ and mid-to-low back vowel between /ɔ/ and /ɑ/ (a retracted version of /a/, as in 'father'). The most frequently occurring 5-vowel inventory (e.g. Spanish) is /i e (or ɛ) a o (or ɔ) u/ (Crothers 1978; Maddieson 1984). The predicted system deviates somewhat from the commonly observed system, especially with respect to the position of the mid-back vowel, but the overall fit is still reasonably good.

The predicted 7-vowel system in figure 6 resembles the predicted 5-vowel system, except for the addition of two new high vowels between /i/ and /u/. These additional high vowels are the back, non-lip-rounded /ɨ/ and the front lip-rounded /ü/ (equivalent to /y/). However, the most commonly observed 7-vowel system (e.g. Italian) is /i e ɛ a ɔ o u/, which includes pairs of mid-front and mid-back vowels but no high vowels between /i/ and /u/.

Thus, the dispersion theory of Liljencrants & Lindblom (1972) performs as good as quantal theory in predicting the frequent occurrence of /i/, /a/, /u/ and /o/ (or /ɔ/), and it outperforms quantal theory in correctly predicting that /e/ (or /ɛ/) will occur frequently and that /y/ will be less common than /i/. The most important failure of the Liljencrants & Lindblom simulations is the prediction of too many high vowels for inventories of seven or more vowels.

Similar to quantal theory, dispersion theory includes both listener- and talker-oriented selection criteria. In the simulations of Liljencrants & Lindblom (1972), the talker-oriented selection criteria were implemented by restricting the available phonetic space (i.e. the output of the vowel production model of Lindblom & Sundberg (1971)) to certain 'basic' articulatory types

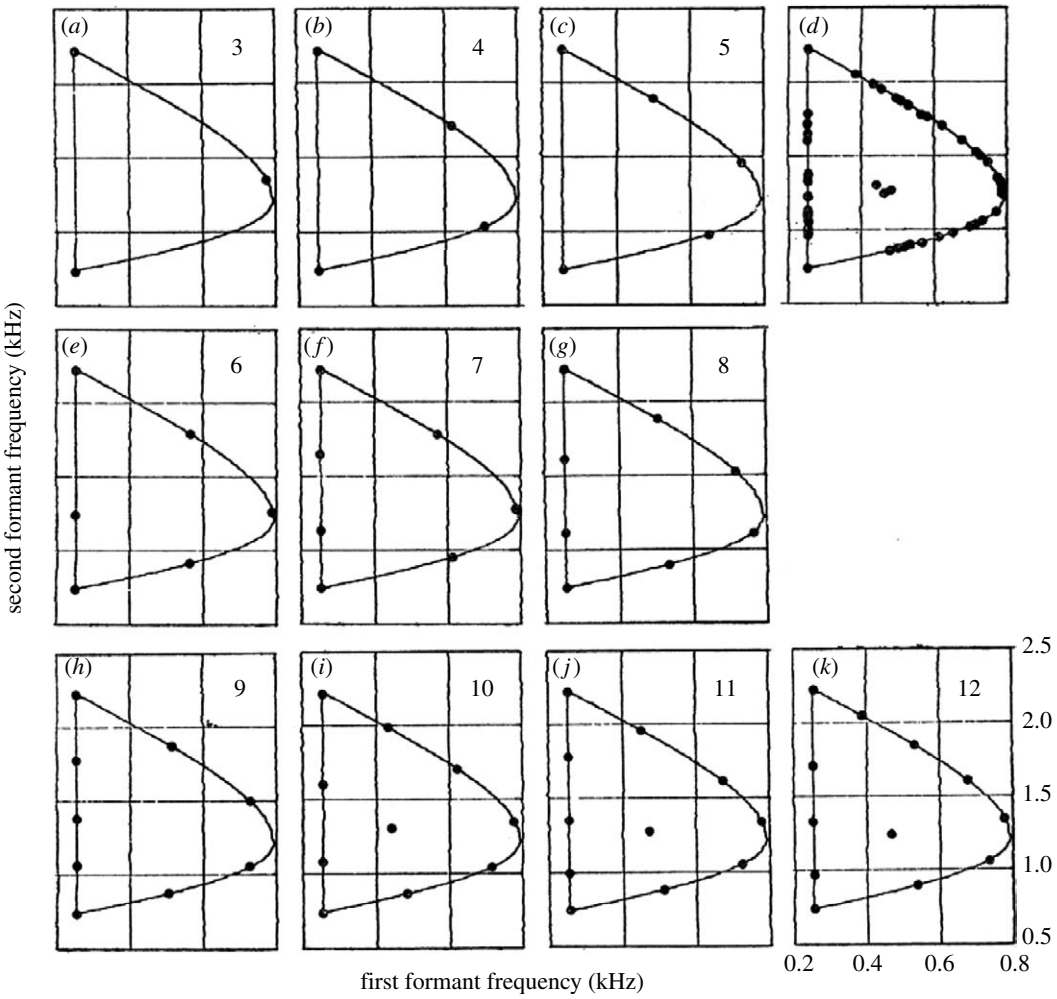

Figure 6. Results of simulations by Liljencrants & Lindblom (1972) of preferred vowel systems ranging in size from (*a–c*,*e–k*) 3 to 12. Solid curves represent the range of possible outputs from the articulatory model of Lindblom & Sundberg (1971), and filled circles correspond to the vowels selected according to a maximum distance criterion. (*d*) Pools all vowels selected across the 10 inventory sizes. (Adapted with permission of the second author from Liljencrants & Lindblom (1972). Linguistic Society of America.)

(e.g. non-nasalized vowels). Later versions of dispersion theory (Lindblom 1986, 1990) attempted to account not only for the structure of preferred vowel inventories and but also for variation in speech clarity in running speech. In these versions of the theory, the goal of talkers is to produce speech that is intelligible to listeners but to do so with as little effort as necessary. In other words, talkers try to achieve sufficient, rather than maximal, distinctiveness, and they thus tend to vary their utterances from reduced ('hypo-speech') forms to clear ('hyper-speech') forms depending on the communication conditions that apply. In general, as information content increases, clarity of speech production also increases (for a review, see Hay *et al.* (2006)). Since the focus of the present paper is on the structure of preferred sound inventories rather than on phonetic variation in running speech, the role of talker-oriented selection factors will not be further discussed.

(e) *Auditory enhancement hypothesis*

It is useful to consider in more detail how talkers implement a listener-oriented strategy of vowel dispersion. One simple approach is to select relatively extreme tongue body and jaw positions since acoustic distinctiveness tends to be correlated with articulatory distinctiveness. As was discussed earlier, the vowels /i/, /a/ and /u/ are each characterized by articulatory extremes in this sense. However, the acoustic dispersion of these vowels is only partly explained by the positioning of the tongue body and jaw. A fuller account is provided by the auditory enhancement hypothesis (Diehl & Kluender 1989*a*,*b*; Diehl *et al.* 1990; Kingston & Diehl 1994), which attempts to explain common patterns of phonetic covariation on listener-oriented grounds. The hypothesis states that phonetic properties of sound categories covary as they do largely because language communities tend to select properties that have mutually enhancing auditory effects. In the case of vowels, auditory enhancement is most typically achieved by combining articulatory properties that have similar—and hence reinforcing—acoustic consequences. (The auditory enhancement hypothesis is closely related to the theory of redundant features independently developed by Stevens and his colleagues (Stevens *et al.* 1986; Stevens & Keyser 1989).)

The high back vowel /u/ offers a good example of how auditory enhancement works. In figure 7 it may be seen that /u/ is distinguished from lower vowels in having a low F1 and from more anterior vowels in

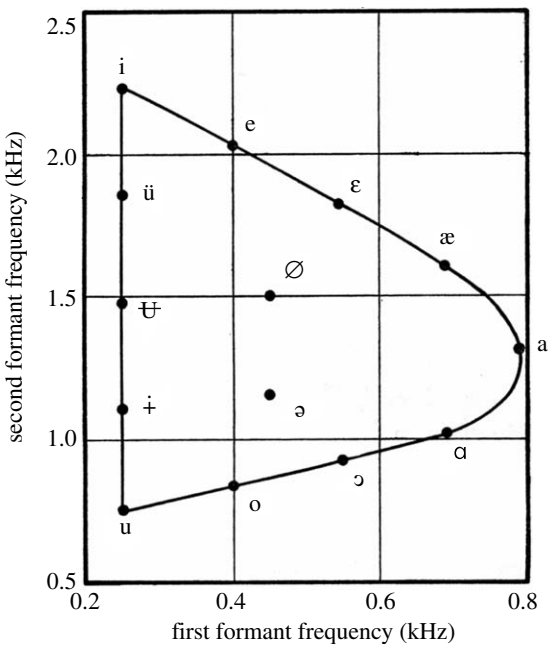

Figure 7. Approximate locations of major vowel categories within the space of outputs from the articulatory model of Lindblom & Sundberg (1971). (Adapted with permission of the second author from Liljencrants & Lindblom (1972), Linguistic Society of America.)

having a low F2. Articulatory properties that produce a lowering of F1 and F2 thus contribute to the distinctiveness of /u/. From acoustic theory (e.g. Fant 1960), it is known that for a tube-like configuration such as the vocal tract, there are several ways to lower a resonance frequency. These are: (i) to lengthen the tube at either end, (ii) to constrict the tube at any node in the standing pressure wave corresponding to the resonance (see earlier discussion), and (iii) to dilate the tube at any antinode in the same standing pressure wave. It happens that in carefully articulated versions of /u/, every one of these options is exploited.

For the purpose of this analysis, the vocal tract is treated in its initial configuration as a quarter-wave resonator, which is then subjected to such perturbations as (i)–(iii) to achieve a target configuration (*viz.* a sequence of two Helmholtz resonators) corresponding to /u/. VT lengthening can be achieved by protruding the lips, a typical component of the lip-rounding gesture that accompanies the production of /u/. Lengthening is also achieved by larynx lowering, and this too has been observed during /u/ production (MacNeilage 1969; Riordan 1977). Both F1 and F2 are lowered as a result of these gestures. The two major VT constrictions characteristic of /u/ occur at the lips (contraction of the lip orifice is another component of the rounding gesture) and in the velar region near the junction between the pharyngeal and oral cavities. As the lip orifice is located at nodes in the standing pressure waves corresponding to the first and second resonances, the effect of lip constriction is to lower F1 and F2. The constriction in the velar region, which is located near another node in the standing pressure wave pattern for the second resonance, yields additional lowering of F2. Finally, VT dilations near the mid-palate and in the lower pharynx (both corresponding to second resonance antinodes) contribute to yet more F2 lowering. The dilation near the mid-palate may be viewed as a by-product of tongue-body retraction and raising, but the pharyngeal dilation is largely the result of tongue-root advancement, which appears to be, anatomically speaking, partly independent of tongue height (Lindau 1979). In summary, the VT configuration for /u/ is optimally tailored to yield an acoustically distinctive vowel. Analogous arguments may be made with respect to other commonly occurring vowels such as /i/ and /a/.

(f) *Attempts to unify dispersion theory and (aspects of) quantal theory*

Following up on the work of Lindblom and his colleagues (Liljencrants & Lindblom 1972; Lindblom 1986) and Stevens (1972, 1989), Schwartz *et al.* (1997) proposed the dispersion–focalization theory of vowel systems. The theory attempts to predict favoured vowel inventories by summing the effects of two perceptual components: global dispersion and local focalization. The first component is based on distances among vowels in a formant frequency space similar to that used by Liljencrants & Lindblom. The second component is based on spectral salience of individual vowels (i.e. the presence or absence of relatively intense spectral regions) and is related to proximity between formants, thus giving weight to an important property of quantal vowels. Simulations of preferred vowel systems are controlled by two free parameters: one sets the relative contribution of F1 versus higher formant frequencies in determining inter-vowel distances and the other sets the relative contributions of the dispersion and the focalization components. Schwartz *et al.* identified a range of values of these parameters that provided reasonably good fits to the structure of the most common vowel inventories. In particular, by giving F1 more weight than higher formant frequencies, the tendency to predict the occurrence of too many high vowels was eliminated (cf. Liljencrants & Lindblom 1972). This outcome is expected because F2 and F3 are primarily related to lip configuration and the front–back position of the tongue, whereas F1 is primarily related to tongue and jaw height. Accordingly, reducing the weight of higher formants effectively compresses the auditory extent of the front–back dimension relative to the height dimension, allowing less room for high vowels such as /ɨ/ and /y/.

Is it possible to approximate the success of the dispersion–focalization theory without using free parameters to obtain an acceptable fit to the data? One approach is to try to improve on formant-based measures of inter-vowel distance by taking into account aspects of the auditory representation of speech signals. For example, Lindblom (1986) adopted a measure of auditory distance based on auditory representations of whole spectra rather than formant frequencies. The representations are derived from a model incorporating auditory filtering (see Moore 2008) as well as pitch and loudness scaling. Auditory distance is defined as the Euclidean distance between two vowels (with the same F0) in an n-dimensional space, where n is the number of auditory filters, and the value for a given vowel on any dimension is equal to the output of the

corresponding filter (scaled in Sones/Bark).[7] When vowel system simulations were carried out using this measure of auditory distance (along with the same inter-vowel distance maximization criterion applied in earlier simulations), the results were disappointing. Although there was some small improvement in predictive accuracy relative to the simulations of Liljencrants & Lindblom (1972), the problem of too many high vowels remained.

The auditory filter outputs used in the simulations by Lindblom (1986) are whole-spectrum representations intended to model (albeit very roughly) the average firing rate of auditory neurons as a function of their best frequency (see Young 2008). Such representations are incomplete in one important respect, namely, temporal information about stimulus frequency (phase locking) is not included. This omission may be significant because such temporal information tends to be more resistant to noise degradation than information contained in average firing rates alone (Sachs *et al.* 1982; Greenberg 1988; Young 2008). In particular, spectral peaks (e.g. formants) are temporally coded by neurons not only with best frequencies closest to the peaks but also with somewhat different best frequencies (Delgutte & Kiang 1984). Thus, temporal coding yields a redundant and fairly noise resistant representation of prominent regions in the stimulus spectrum. Since normal speech communication takes place in the presence of background noise, a measure of auditory distance that ignores temporal coding may yield inaccurate predictions about preferred speech sound inventories.

With these considerations in mind, Diehl *et al.* (2003) conducted a new series of vowel system simulations with an auditory model that incorporates an analogue of average firing rate as a function of best frequency as well as a dominant frequency representation based on temporal coding. For any two vowels with the same F0, these two forms of representation are multiplied to form a single spatio-temporal measure of auditory distance.[8] The vowel systems predicted by these simulations show a reasonably good fit to the most commonly occurring systems. For example, the predicted 7-vowel system includes /i/, /a/ and /u/ as well as two mid-front vowels and two mid-back vowels, similar to Italian and most other 7-vowel systems. In other words, the problem of too many high vowels (Liljencrants & Lindblom 1972; Lindblom 1986) is eliminated.

Why does inclusion of temporal coding information improve the accuracy of vowel system simulations? Owing to redundant specification of relatively intense frequency components, formant peaks contribute disproportionately to auditory representations and hence to measures of auditory distance. The first formant plays an especially large role owing to its greater intensity relative to higher formants. This produces a perceptual warping of the vowel space such that the vowel height dimension (corresponding to F1) can perceptually accommodate more vowel contrasts than the front–back dimension (corresponding to F2 and F3), and this in turn reduces the likelihood that high vowels between /i/ and /u/ will be selected by the maximum distance criterion. Temporal coding also boosts the contrastive value of quantal vowels, since their closely spaced formants give rise to spectrally salient regions that are redundantly specified in the auditory representation.

5. CONCLUDING REMARKS

The modelling approach of Diehl *et al.* (2003) yields results that are generally similar to those of the dispersion–focalization theory (Schwartz *et al.* 1997). However, Diehl *et al.* do not treat global dispersion and local focalization (spectral salience) as separate perceptual components. Instead, spectral salience directly enhances global dispersion as an automatic consequence of spatio-temporal coding of frequency. In this way, key elements of dispersion and quantal theories are unified within a single explanatory framework.

I thank Björn Lindblom for many years of productive discussion of these issues. I am also grateful to Andrew Lotto, Jessica Hay, Sarah Sullivan, Brian Moore, Thomas Baer and Christopher Darwin for their very helpful comments on an earlier draft of this paper and to Sarah Sullivan for her help in preparing the figures. Preparation of this paper was supported by NIH grant R01 DC000427-15.

ENDNOTES

[1]Catford (1977) distinguishes between two types of frication source: 'channel turbulence', which is produced simply by airflow through a channel, and 'wake turbulence', which is created downstream from the edge of an obstacle (e.g. teeth or upper lip) oriented perpendicular to the airflow.

[2]In addition to resonances, a VT filter function may be characterized by antiresonances, which have the opposite effect on the spectrum. At or near the frequency of an antiresonance, energy from a source is absorbed and hence greatly attenuated in the output spectrum. Antiresonances are introduced into the filter function if (i) the vocal tract has a side branch or bifurcated airways, as in the production of nasal consonants or nasalized vowels or (ii) there is an occlusion or narrow constriction of the vocal tract, as in the production of stop or fricative consonants (Kent & Read 1992).

[3]Tom Baer's helpful suggestions about the wording of this paragraph are gratefully acknowledged.

[4]A more realistic model for /i/ would include a third tube at the front of the vocal tract larger in diameter than the second tube. This extended model incorporates the effects of lip spreading (Stevens 1998).

[5]Consonants tend to be briefer in duration and less intense than vowels, but their perception is generally robust. As discussed in §4*c*, quantal properties of certain consonants help to explain this robustness. However, another important factor is the dynamic character of consonant production, which gives rise to a rich set of time-distributed perceptual cues. For example, the identity of a word-medial stop is signalled by properties of the vowel–consonant transitions, the occlusion interval (e.g. its duration), the transient burst of energy at the articulatory release, the fricative and/or aspiration following the burst and the consonant–vowel transitions (Pickett 1999). For further discussion of cue redundancy in consonant perception, see Kingston & Diehl (1994).

[6]This notion of maximal dispersion is apparently what earlier linguists such as Passy (1890), Jakobson (1941) and Martinet (1955) had in mind when they discussed the role of perceptual contrast in the structure of speech sound systems.

[7]Sones are units of subjective loudness; Barks are units of subjective frequency, with one Bark corresponding to a step in frequency equal to one critical band (Zwicker & Terhardt 1980; also see Moore 2008).

[8]To model temporal coding, an inverse FFT is performed on the spectral output of each auditory filter and the resulting time-domain signal is input to a dominant frequency detector, which specifies dominant frequency in terms of zero crossings (Carlson & Granström

1982). The output of these detectors (one per auditory filter) is the dominant frequency representation for a given vowel. To calculate auditory distance between two vowels, the product of the average firing rate representation and the dominant frequency representation is computed for each filter, and the Euclidean distances are then calculated as in Lindblom (1986).

REFERENCES

Abramson, A. S. & Lisker, L. 1970 Discriminability along the voicing continuum: cross-language tests. In *Proc. 6th Int. Cong. of Phonetic Sciences, Prague, 1967*, pp. 569–573. Prague, Czech Republic: Academia.

Aslin, R. N., Pisoni, D. B., Hennessy, B. L. & Perey, A. J. 1981 Discrimination of voice onset time by human infants: new findings and implications for the effects of early experience. *Child Dev.* **52**, 1135–1145. (doi:10.2307/1129499)

Carlson, R. & Granström, B. 1982 Towards an auditory spectrograph. In *The representation of speech in the peripheral auditory system* (eds R. Carlson & B. Granström), pp. 109–114. Amsterdam, The Netherlands: Elsevier Biomedical.

Catford, J. C. 1977 *Fundamental problems in phonetics*. Bloomington, IN: Indiana University Press.

Chiba, T. & Kajiyama, M. 1941 *The vowel: its nature and structure*. Tokyo, Japan: Tokyo-Kaisekan. (Reprinted by the Phonetic Society of Japan 1958.)

Chomsky, N. & Halle, M. 1968 *The sound pattern of English*. New York, NY: Harper & Row.

Crothers, J. 1978 Typology and universals of vowel systems. In *Universals of human language*, vol. 2 (eds J. H. Greenberg, C. A. Ferguson & E. A. Moravcsik), pp. 99–152. Stanford, CA: Stanford University Press.

Darwin, C. J. 2008 Listening to speech in the presence of other sounds. *Phil. Trans. R. Soc. B* **363**, 1011–1021. (doi:10.1098/rstb.2007.2156)

Delgutte, B. & Kiang, N. Y.-S. 1984 Speech coding in the auditory nerve I: vowel-like sounds. *J. Acoust. Soc. Am.* **75**, 866–878. (doi:10.1121/1.390596)

Diehl, R. L. 1989 Remarks on Stevens' quantal theory of speech. *J. Phonet.* **17**, 71–78.

Diehl, R. L. & Kluender, K. R. 1989*a* On the objects of speech perception. *Ecol. Psychol.* **1**, 121–144. (doi:10.1207/s15326969eco0102_2)

Diehl, R. L. & Kluender, K. R. 1989*b* Reply to commentators. *Ecol. Psychol.* **1**, 195–225. (doi:10.1207/s15326969eco0102_6)

Diehl, R. L. & Lindblom, B. 2004 Explaining the structure of feature and phoneme inventories: the role of auditory distinctiveness. In *Speech processing in the auditory system* (eds S. Greenberg, W. A. Ainsworth, A. N. Popper & R. R. Fay), pp. 101–162. New York, NY: Springer.

Diehl, R. L., Kluender, K. R. & Walsh, M. A. 1990 Some auditory bases of speech perception and production. In *Advances in speech, hearing and language processing*, vol. 1 (ed. W. A. Ainsworth), pp. 243–268. London, UK: JAI Press.

Diehl, R. L., Lindblom, B. & Creeger, C. P. 2003 Increasing realism of auditory representations yields further insights into vowel phonetics. In *Proc. 15th Int. Cong. of Phonetic Sciences*, vol. 2, pp. 1381–1384. Adelaide, Australia: Causal Publications.

Diehl, R. L., Lotto, A. J. & Holt, L. L. 2004 Speech perception. *Annu. Rev. Psychol.* **55**, 149–179. (doi:10.1146/annurev.psych.55.090902.142028)

Fant, G. 1960 *Acoustic theory of speech production*. The Hague, The Netherlands: Mouton.

Fant, G. 1973 *Speech sounds and features*. Cambridge, MA: MIT Press.

Flanagan, J. L. 1972 *Speech analysis synthesis and perception*. Berlin, Germany: Springer.

Greenberg, S. 1988 Acoustic transduction in the auditory periphery. *J. Phonet.* **16**, 3–17.

Hay, J. F., Sato, M., Coren, A. E., Moran, C. L. & Diehl, R. L. 2006 Enhanced contrast for vowels in utterance focus: a cross-language study. *J. Acoust. Soc. Am.* **119**, 3022–3033. (doi:10.1121/1.2184226)

Jakobson, R. 1941 *Kindersprache, Aphasie und allgemeine Lautgesetze*, pp. 1–83. Uppsala, Sweden: Uppsala Universitets Arsskrift.

Jakobson, R., Fant, G. & Halle, M. 1963 *Preliminaries to speech analysis*. Cambridge, MA: MIT Press.

Johnson, K. 1997 *Acoustic and auditory phonetics*. Cambridge, MA: Blackwell.

Kent, R. D. & Read, C. 1992 *The acoustic analysis of speech*. San Diego, CA: Singular Publishing.

Kingston, J. & Diehl, R. L. 1994 Phonetic knowledge. *Language* **70**, 419–454. (doi:10.2307/416481)

Klatt, D. H. 1982 Prediction of perceived phonetic distance from critical-band spectra: a first step. In *Proc. IEEE Int. Conf. Speech, Acoustic Signal Process*, vol. 82, pp. 1278–1281.

Kuhl, P. K. & Miller, J. D. 1978 Speech perception by the chinchilla: identification functions for synthetic VOT stimuli. *J. Acoust. Soc. Am.* **63**, 905–917. (doi:10.1121/1.381770)

Ladefoged, P., Harshman, R., Goldstein, L. & Rice, L. 1977 Vowel articulation and formant frequencies. *UCLA Working Papers Phonet.* **38**, 16–40.

Lasky, R. E., Syrdal-Lasky, A. & Klein, R. E. 1975 VOT discrimination by four to six and a half month old infants from Spanish environments. *J. Exp. Chid Psychol.* **20**, 215–225. (doi:10.1016/0022-0965(75)90099-5)

Liljencrants, J. & Lindblom, B. 1972 Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language* **48**, 839–862. (doi:10.2307/411991)

Lindau, M. 1979 The feature expanded. *J. Phonet.* **7**, 163–176.

Lindblom, B. 1986 Phonetic universals in vowel systems. In *Experimental phonology* (eds J. J. Ohala & J. J. Jaeger), pp. 13–44. Orlando, FL: Academic Press.

Lindblom, B. 1990 Explaining phonetic variation: a sketch of the H & H theory. In *Speech production and speech modeling* (eds W. J. Hardcastle & A. Marchal), pp. 403–439. Dordrecht, The Netherlands: Kluwer.

Lindblom, B. & Sundberg, J. 1971 Acoustical consequences of lip, tongue, jaw and larynx movement. *J. Acoust. Soc. Am.* **50**, 1166–1179. (doi:10.1121/1.1912750)

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

Lisker, L. & Abramson, A. S. 1970 The voicing dimension: some experiments in comparative phonetics. In *Proc. 6th Int. Cong. of Phonetic Sciences, Prague 1967*, pp. 563–567. Prague, Czech Republic: Academia.

MacNeilage, P. M. 1969 A note on the relation between tongue elevation and glottal elevation. Monthly Internal Memorandum, University of California, Berkeley, January 1969, pp. 9–26.

Maddieson, I. 1984 *Patterns of sound*. Cambridge, UK: Cambridge University Press.

Martinet, A. 1955 *Économie des Changements Phonétiques*. Berne, Switzerland: Francke.

Miller, J. D., Wier, C. C., Pastore, R. E., Kelly, W. J. & Dooling, R. J. 1976 Discrimination and labeling of noise–buzz sequences with varying noise-lead times: an example of categorical perception. *J. Acoust. Soc. Am.* **60**, 410–417. (doi:10.1121/1.381097)

Moore, B. C. J. 2008 Basic auditory processes involved in the analysis of speech sounds. *Phil. Trans. R. Soc. B* **363**, 947–963. (doi:10.1098/rstb.2007.2152)

Passy, P. 1890 *Études sur les Changements Phonétiques et Leurs Caractères Généraux*. Paris, France: Librairie Firmin-Didot.

Perkell, J. S. 1979 On the nature of distinctive features: implications of a preliminary vowel production study. In *Frontiers of speech communication research* (eds B. Lindblom & S. Őhman), pp. 365–380. New York, NY: Academic Press.

Pickett, J. M. 1999 *The acoustics of speech communication*. Needham Heights, MA: Allyn & Bacon.

Pisoni, D. B. 1977 Identification and discrimination of the relative onset time of two component tones: implications for voicing perception in stops. *J. Acoust. Soc. Am.* **61**, 1352–1361. (doi:10.1121/1.381409)

Pisoni, D. B. 1980 Variability of vowel formant frequencies and the quantal theory of speech: a first report. *Phonetica* **37**, 285–305.

Repp, B. H. 1984 Categorical perception: issues, methods, findings. In *Speech and language: advances in basic research and practice*, vol. 10 (ed. N. J. Lass), pp. 243–335. New York, NY: Academic Press.

Riordan, C. J. 1977 Control of vocal-tract length in speech. *J. Acoust. Soc. Am.* **62**, 998–1002. (doi:10.1121/1.381595)

Sachs, M., Young, E. & Miller, M. 1982 Encoding of speech features in the auditory nerve. In *The representation of speech in the peripheral auditory system* (eds R. Carlson & B. Granström), pp. 115–130. Amsterdam, The Netherlands: Elsevier Biomedical.

Schwartz, J.-L., Boë, L.-J., Vallée, N. & Abry, C. 1997 The dispersion–focalization theory of vowel systems. *J. Phonet.* **25**, 255–286. (doi:10.1006/jpho.1997.0043)

Sinex, D. G., McDonald, L. P. & Mott, J. B. 1991 Neural correlates of nonmonotonic temporal acuity for voice onset time. *J. Acoust. Soc. Am.* **90**, 2441–2449. (doi:10.1121/1.402048)

Stevens, K. N. 1972 The quantal nature of speech: evidence from articulatory–acoustic data. In *Human communication: a unified view* (eds E. E. David & P. B. Denes), pp. 51–66. New York, NY: McGraw-Hill.

Stevens, K. N. 1989 On the quantal nature of speech. *J. Phonet.* **17**, 3–45.

Stevens, K. N. 1998 *Acoustic phonetics*. Cambridge, MA: MIT Press.

Stevens, K. N. & House, A. S. 1955 Development of a quantitative description of vowel articulation. *J. Acoust. Soc. Am.* **27**, 484–493. (doi:10.1121/1.1907943)

Stevens, K. N. & House, A. S. 1961 An acoustical theory of vowel production and some of its implications. *J. Speech Hear. Res.* **4**, 303–320.

Stevens, K. N. & Keyser, S. J. 1989 Primary features and their enhancement in consonants. *Language* **65**, 81–106. (doi:10.2307/414843)

Stevens, K. N., Keyser, S. J. & Kawasaki, H. 1986 Toward a phonetic and phonological theory of redundant features. In *Invariance and variability in speech processes* (eds J. S. Perkell & D. H. Klatt), pp. 426–449. Hillsdale, NJ: Erlbaum.

Syrdal, A. K. & Gopal, H. S. 1986 A perceptual model of vowel recognition based on the auditory representation of American English vowels. *J. Acoust. Soc. Am.* **79**, 1086–1100. (doi:10.1121/1.393381)

Young, E. D. 2008 Neural representation of spectral and temporal information in speech. *Phil. Trans. R. Soc. B* **363**, 923–945. (doi:10.1098/rstb.2007.2151)

Zipf, G. K. 1949 *Human behavior and the principle of least effort: an introduction to human ecology*. Cambridge, MA: Addison-Wesley.

Zwicker, E. & Terhardt, E. 1980 Analytical expressions for critical band rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am.* **68**, 1523–1525. (doi:10.1121/1.385079)