after speech has been initiated, the amount of time to pause for silence detection can be increased.

Conclusion

Designing an effective speech user interface is an iterative process. Not even the most experienced designer will craft a perfect dialogue in the first iteration. To ensure that the system will be used over time, focus on users and do not be afraid to modify the design as new data are collected. Observe users during the task modeling phase to understand who they are and what are their goals. Listen carefully to users while conducting natural dialogue studies to determine how they speak in the context of the task. Test the design with target users to ensure that the prompts are clear, that feedback is appropriate, and that errors are caught and corrected. In addition, when testing, verify that the design is accomplishing the business goals that were set out to be achieved. If problems are uncovered during the test, revise the design and test again. By focusing on users and iterating on the design, one can produce an effective, polished speech interface design.

See also: Speech Recognition: Statistical Methods; Speech Synthesis.

Bibliography

Baecker R & Buxton W (eds.) (1987). *Readings in human-computer interaction: a multidisciplinary approach.* Los Altos, CA: Morgan-Kaufmann Publishers.

Clark H (1993). *Arenas of language use.* Chicago: University of Chicago Press.

Cohen M, Giangola J & Balogh J (2004). *Voice user interface design.* Boston: Addison-Wesley.

Cooper A (1995). *About face: the essentials of user interface design.* Foster City, CA: IDG Books.

Java Speech Grammar Format (1998). 'Specification Version 1.0.' http://java.sun.com/products/java-media/speech/forDevelopers/JSGF.

Lai J & Yankelovich N (2003). 'Conversational speech interfaces.' In Jacko J & Sears A (eds.) *The human–computer interaction handbook: fundamentals, evolving technologies and emerging applications.* Mahwah, NJ: Lawrence Erlbaum. 698–713.

Nielsen J (1993). *Usability engineering.* Boston: Academic Press.

Preece J, Rogers Y & Sharp H (2002). *Interaction design: beyond human–computer interaction.* New York: John Wiley.

Reeves B & Nass C (1996). *The media equation: how people treat computers, television and new media like real people and places.* New York: Cambridge University Press/CSLI.

Speech Perception

H Mitterer and A Cutler, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Introduction

When we listen to speech, our aim is to determine the meaning of the acoustic input. Only very rarely is a listener – usually some kind of speech scientist – concerned with the sounds of speech instead of with the content. Yet the study of speech perception has largely dealt with how listeners tell phonemes – the sounds of speech – from one another. Why is this so?

Phonemes are, by definition, the minimal units that distinguish between words, i.e., between one meaning and another. Listeners may know tens, indeed hundreds of thousands of words, and may still be able to learn new words effortlessly on a daily basis. However, languages construct this vast stock of words from, on average, only around 30 separate phonemes (Maddieson, 1984; see Figure 1). Thus, there is both parsimony and validity in the speech perception research program. Parsimony, because study of the perceptual cues to a few dozen phonemes is tractable in a way that study of the perceptual cues to individual words is not, and validity, because understanding how listeners tell, say, a /b/ from a /p/ informs us about the recognition of all word pairs that differ in these phonemes. It is important to realize that this research focus on cues to phonemic identity did not proceed from an assumption that phonemes necessarily played a role as units of speech perception. In fact, there have been many competing proposals concerning the basic unit of speech perception, all more or less unsatisfactory (Klatt, 1989), from units below the level of the phoneme (phonological features) to above it (diphones, syllables) to no sublexical unit at all. Rather, the phonemic research focus was the result of rational task analysis: to understand speech, listeners need to recognize words, and the process of recognizing words involves distinguishing a word from all other words that are its closest neighbors in the vocabulary – *word* from *bird*, *ward*, *work*, and so on. By making such distinctions, the listener

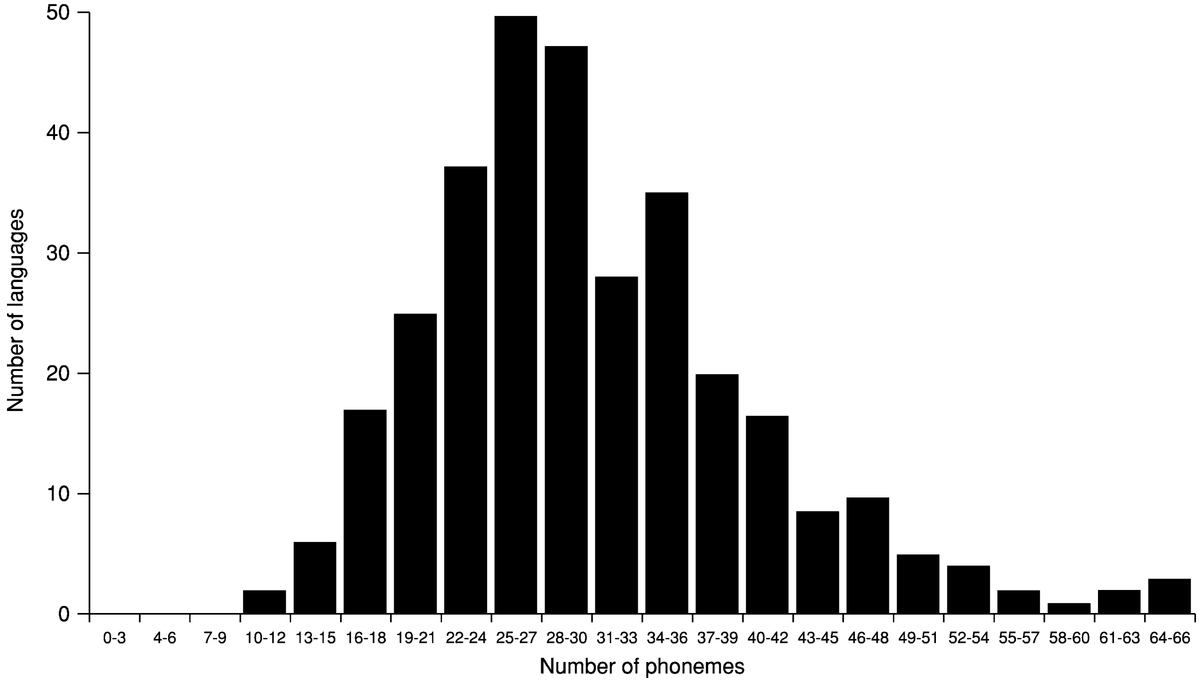

Figure 1 Distribution of phoneme inventory size across the representative sample of 317 languages analyzed by Maddieson (1984). The average number of phonemes is 31.

is effectively making decisions that are phonemic in nature.

True to the history of speech perception research, therefore, this review is organized first by phonemic category and the type of information that enables listeners to recognize members of each category; following sections deal with the principal theories that have driven speech perception research; and we conclude with a section on how this body of research relates to the recognition of the larger, meaningful units that really concern listeners.

Phoneme Perception

Phonemes in Real Speech

If words can indeed be viewed as a sequence of phonemes, it might seem reasonable to assume that these sounds are concatenated one after another in the speech stream. This is far from the case. Speech signals are continuous and it is hard to discern where one word ends and the next begins, let alone locate the borders of individual phonemes (see Figure 2). An analogy concocted by Hockett (1955: 210) still appears in most introductory textbooks: phonemes in speech are like colored Easter eggs moving on a conveyor belt to be crushed by a wringer. The wringer smears out the eggs so that no part of the conveyor belt is exclusively covered by the remains of one particular Easter egg. As a consequence, the coloring of one egg is mixed with the colorings of the surrounding eggs. This analogy is intended to convey that the speech signal is neither separable nor invariant. It is not separable because no part of the speech signal is exclusively influenced by one phonological unit. It is not invariant because adjacent – in fact even nonadjacent – phonemes are coarticulated, so that acoustic correlates of a given phoneme vary. Therefore, running speech is completely different from, for instance, Morse code, in which invariant acoustic signals are concatenated. In making this important point, the lack-of-invariance argument may sometimes be exaggerated. Some phoneme classes – for example, voiceless fricatives – can have quite salient and local cues (see the different tokens of /s/ in Figure 2). Others, however, show a great deal of variation. Stop consonants are a case in point; exactly the same stretch of speech can be heard as /p/ before one vowel, but as /k/ before another (see Figure 3). Vowels can also be highly coarticulated; thus the second formant (F2) in a vowel will generally be higher after alveolar consonants than after labial consonants, corresponding to the higher F2 locus in alveolar than in labial consonants.

Sources of variance arise both within and between speakers. Not only the phonetic context in which a phoneme occurs, but also the rate of speech and the speech register (from formal and careful to casual and careless) can affect how that phoneme is realized

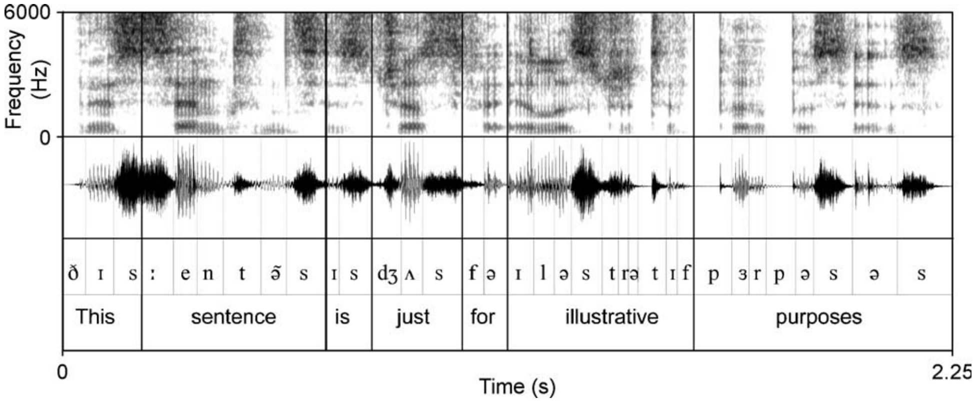

Figure 2 An oscillogram and spectrogram of the sentence *This sentence is just for illustrative purpose*s spoken by an adult male native speaker of British English born in Scotland.

within the utterances of a given speaker. So too can the degree of prosodic prominence within the sentence of the word in which the phoneme occurs. Variance between speakers arises most noticeably from physiological differences, e.g., in that the formant frequencies of vowels reflect the resonances of the vocal tract. The resonances are lower the larger the vocal tract; thus children produce higher formant frequencies than adults, and adults also differ. Further interspeaker variance is introduced by dialectal variation.

Variability may sometimes be so extensive that, superficially, a phonemic contrast is blurred. For instance, most languages that use nasals in word-final position allow the place of articulation to be assimilated to the place of a following consonant; the final consonant of *sun* may be pronounced /n/ in *suntan*, /m/ in *sunbathing*, /ng/ in *sunglasses*. In running speech, sounds may be deleted (e.g., the word *just* in Figure 2 contains no [t]) or inserted (an utterance of *something* may contain cues consistent with a /p/ between the two syllables, or an utterance of *pensive* may contain cues consistent with a /t/).

All of this variability makes the perception of cues to phonemes a nontrivial task for listeners. A large body of early speech research was dedicated to the search for invariant acoustic properties of phonemes. This enterprise may be reckoned unsuccessful with respect to its ultimate goal. Nevertheless, these efforts have certainly succeeded in finding acoustic properties, or cues, that correlate with certain phonemes or phoneme classes. This body of work represents a cornerstone of the study of speech perception and describes the basic cues for the recognition of phoneme classes and distinction within these classes.

Vowels Vowels are usually the parts of the speech signal containing local maximal amplitudes and periodicity caused by the vibration of the vocal folds. With these two cues, vowels can be reasonably well distinguished from consonants (see Figure 2). The vocal fold vibration gives rise to a periodic source signal with a large number of harmonics. This source signal is then filtered by the vocal tract (*see* **Speech Production**). The vocal tract amplifies some of the harmonics due to its resonance characteristics. Regions with amplified harmonics are called formants. The frequencies of these formants depend on the exact shape of the vocal tract, that is, on tongue position and shape, the position of the jaws, etc., (*see* **Speech Production**). Accordingly, vowels can be distinguished from one another by their steady state formant frequencies. The vowel system of a particular language is often presented in a two-dimensional vowel space with first-formant frequency on the ordinate and the second-formant frequency on the abscissa. This representation gives rise to a vowel triangle with the vowels [u] as in *shoe*, [i] as in *she*, and [a] as in *shah* as corners. Other vowels can be associated with different positions in this vowel triangle. Accordingly, listeners can identify vowels by computing the position in formant space (Nearey, 1989).

A number of facts complicate this simple picture. First, some languages also use diphthongal vowels, which are characterized by a formant movement throughout the vowel (as the vowel in *shy*). These diphthongs are usually contrasted to steady-state monophthongs, though categorization of vowels as either diphthongal or monophthong is in fact often difficult. Second, cues to vowel identity are also provided by duration – even in languages that do not distinguish long from short vowels – and by dynamic properties of consonant (C)-vowel (V) and VC transitions. Evidence for the importance of transitions comes from studies with so-called silent center CVC syllables, in which the steady-state portion of the vowel has been replaced by silence. Despite this

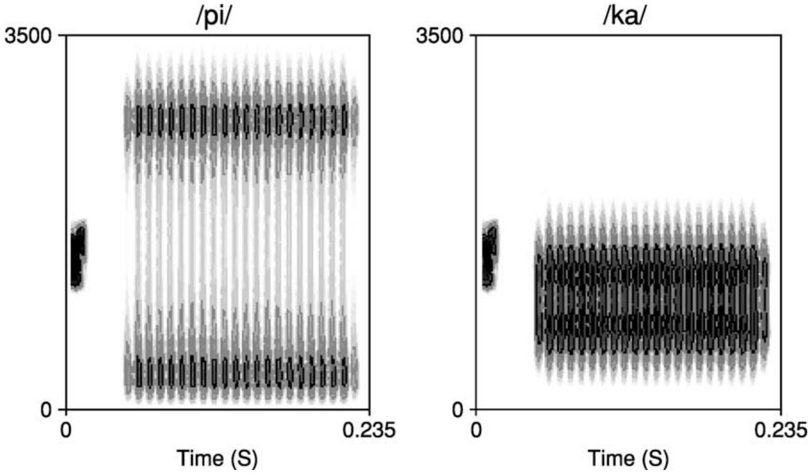

Figure 3 Two synthetic syllables with the same noise burst at the onset as used by Liberman (1996: Chap. 3). The two syllables are perceived as /pi/ and /ka/, that is, with different stop sounds, despite the identical noise burst.

gross mutilation of the vowel, identification accuracy is only slightly impaired. This cannot simply be attributed to a 'good-continuation' interpolation strategy over the silent part of the syllable that effectively restores the steady state, because silent-center syllables are still well identified if the onset and offset stem from different speakers. Instead, the dynamic consonant–vowel properties seem to carry higher-order invariants for vowel identification (Strange, 1999).

Sufficiency of one type of cue of course does not render other types of cue superfluous; it is clear that listeners can exploit both static and dynamic information in vowel perception.

Semivowels and Stops – Dynamic Consonants The classes of semivowels and stop consonants share the property of being defined by articulatory and hence formant movement. The labial stop [b] and labial glide [w] show similar formant transitions. The difference between the two sounds lies in the speed of the formant movement, which is faster for the stop than the semivowel. The semivowel [w] could be described as lying between the vowel [u] and the stop [b]; the stop has a fast rate of formant movement, the vowel has no or only a small amount of inherent movement, while the semivowel has a moderate amount of movement (Liberman, 1996: Chap. 6). Within the stop and semivowel classes, distinctions can be described by the concept of locus (Liberman, 1996: Chap. 5). Thus, [b] and [w] are associated with a labial place of articulation, while the corresponding [d] and [j] are associated with a alveolar place of articulation. Formant movements originate at the appropriate locus for each place or articulation, with especially the second and third formant being relevant for place distinctions. However, the actual onset of the formant frequencies will not be invariant (see Figure 2, with different formant-transition onsets in *purposes* for /pɜ/ and /pə/). The onset of the formant frequency is codetermined by the vowel. This has led to the concept of locus equations, which allow the place of articulation to be inferred based on formant-onset frequency and steady-state vowel frequency (see Sussman *et al.*, 1998, and the associated commentaries for strength and weaknesses of this concept).

Another well-studied distinction is the case of the voicing contrast for stops. English, for instance, contrasts voiced with unvoiced stops, as in *bath* versus *path*. The main cue for this distinction is voice-onset time (VOT). VOT is the temporal disparity between the onset of the opening of the vocal tract and the period movements of the vocal folds. The acoustic cues that arise as a consequence of this timing difference will be, in the case of a negative VOT (the onset of voicing then precedes the opening of the vocal tract), a 'prevoice bar,' a low-frequency hum. In case of a positive VOT, an aperiodic release burst is followed by aspiration. The exact location for which listeners switch from perceiving a voiced to a voiceless stop is language-specific, with some languages making a three-way distinction between prevoiced, unvoiced, and unvoiced aspirated stops.

Nasals and Fricatives – Static Consonants Nasals and fricatives share the attribute of having a stationary part. In Figure 2, which shows an example of running speech, six examples of [s] and one [n] can be seen. In all cases, a short but nevertheless reasonably stationary portion of the speech signal may be observed. Nasals are generally characterized by a low-amplitude, low-frequency voiced signal (caused by an antiresonance; *see* **Speech Production**). Voiceless fricatives, in contrast, contain high-frequency noise. In the case of voiced fricatives, this noise is accompanied by a low-amplitude voicing. Within-class differences are signaled by spectral properties. For fricatives, lower fricative noises signal a more retracted place of articulation.

However, formant transitions in surrounding vowels also contribute to the identification of nasals and fricatives. For the case of prevocalic nasals, formant transitions in the vowel seem to be the most important cue. For postvocalic nasals and fricatives, the formant transitions become less important, while the importance of the steady-state spectral cues increases. The identification of nasals is, however, further complicated by reductions in running speech, such as assimilation, or deletion with nasalization of the previous vowel (as in the second nasal in *sentence* in Figure 2).

Laterals and Rhotics The recognition of laterals, e.g., [l] in the English word *lap*, and rhotics, e.g., [ɹ] in the English word *wrap*, is especially complicated. First of all, implementation of an /l/–/r/ contrast varies strongly across languages. In American English, the contrast is carried mainly by the frequency of the third formant. In Hungarian, the /l/–/r/ contrast is a contrast between a lateral and a trill, which is mainly carried by the presence or absence of amplitude modulation. While laterals can be defined phonetically, rhotics in different languages can be so different that some have retreated to the position that "the terms rhotic and 'r-sound' are largely based on the fact that these sounds tend to be written with a particular character ... namely the letter 'r'" (Ladefoged and Maddieson, 1996: 215). This diversity is countered chiefly by the fact that these sounds seem to replace one another in related languages, or even in different accents of the same language.

In a given language, both /l/ and /r/ often vary in their pronunciation as a function of position in the syllable. Moreover, large interindividual sociophonetic and geographic variations are especially salient for these phonemes. Finally, the acoustic manifestations of the difference between [l] and [r] in English are particularly nonlocal. Reliable acoustic differences between words containing [l] and [r] may extend for more than one syllable to the left and right. As such, the perception of these segments proves to be a challenge for models that assume that spoken language must be recognized via identification of phonemes.

Trading Relations A list of cues to phoneme identity invokes an overly simple picture of speech perception: the perceptual system computes certain values, such as VOT, and then a language-specific mechanism applies simple rules such as (if [VOT > 0.02s] ⇒ voiceless). Alas, the speech signal is both too poor and rich to support such simple decision making. The signal is too poor in that, as described above, it lacks clear invariant properties that are reliably associated with a given speech sound. But secondly, it is also too rich, because a multitude of cues exist for each contrast. Consider the intervocalic voicing contrast in English *rabid* vs. *rapid*. Lisker (1986) lists no fewer than 16 cues to this distinction, including length of the preceding vowel, closure duration, and pitch contour. This redundancy partly compensates for lack of invariance, because information in one dimension can be 'traded' against information in another. Thus, a VOT more consistent with *rapid* may be over-ridden if the length of the preceding vowel strongly cues *rabid*. Note that it is in part due to this redundancy that a reasonably comprehensible signal may remain even after quite extreme deformation of the speech signal, for instance, silencing of the center of vowels.

Categorical Perception

As an empirical phenomenon, categorical perception is probably the most-oft replicated effect in speech perception. Consider the way in which the onset of the second formant crucially distinguishes [da] from [ba]. If in a sequence of synthetic syllables the second formant's onset frequency is varied continuously from a low [ba]-like value to a high [da]-like value, one should expect that listeners would report hearing the syllables becoming progressively more [da]-like. However, this is not what happens. Instead, listeners first report hearing token after token of [ba], and then suddenly change to reporting [da]. Only a small subset of stimuli is perceived ambiguously, that is, sometimes as [ba] and sometimes as [da] (see Figure 4).

In a classic experiment Liberman (1996: Chap. 10) evaluated both the categorization of such a continuum and the discrimination of sounds along this continuum in an ABX task (in which listeners hear three stimuli and decide whether the third is identical to either the first or the second). Listeners showed little ability to discriminate stimuli that were identified as the same phoneme, while they succeeded in between-category discrimination. This finding was of particular interest because it contrasts with the relation of identification and discrimination in other auditory perception domains. In pitch perception, for instance,

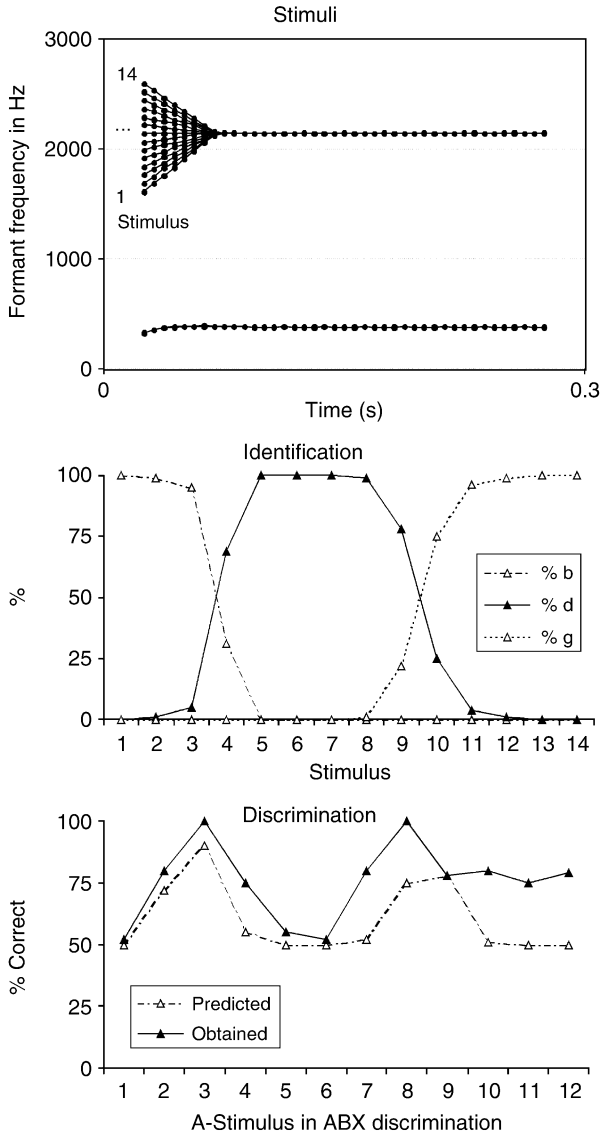

Figure 4 Illustration of categorical perception after Liberman *et al.* (reprinted in Liberman, 1996). The upper panel shows the stimuli 1–14, differing in the second formant transition. The middle panel shows the identification of the stimuli as either /ba/, /da/, or /ga/. The lower panel shows the results of a two-step discrimination task, in which stimulus one had to be discriminated from stimulus three. The figure shows that discrimination peaks coincide with category boundaries, but also that discrimination is generally better predicted by categorization.

discrimination performance exceeds identification performance: listeners can discriminate more than 1000 pitches, but have difficulty forming more than seven categories reliably. Moreover, the categoricality of perception was also found to disappear when formant transitions were presented without a vowel context. Consequently, there is no sharp discontinuity in perception, but each 'chirp' is perceived as equally different from neighboring members of the continuum. This finding further suggested that categorical perception was speech-specific.

Later findings, however, undermined this claim. For VOT contrasts, it has been shown that the boundary effects rest on discontinuities in auditory processing (*see* **Categorical Perception in Animals**). Moreover, the success of within-category discrimination is always better than chance, and more importantly, depends on the task and its implementation. The steepness of identification functions may, hence, be more aptly attributed to categoricalness at a decision rather than a perceptual level (see the chapters in Harnard, 1987).

The steepness of identification functions also varies with the speech contrast tested. Perception of vowels tends to be more continuous and within-category discrimination is easier to elicit. However, discrimination nevertheless seems more difficult if one of the two vowels to be discriminated is close to a prototypical vowel in a given language. This perceptual-magnet effect was supposed to be specific for speech perception (Kuhl, 1991). Later developments, however, cast doubt on the perceptual-magnet effect as such, and its specificity to human speech perception (e.g., Kluender *et al.*, 1998).

Two further findings in the categorical perception tradition indicate that the identification of a given stimulus varies with its environment (see the chapters in Harnad, 1987, for reviews). The first, an effect of the immediate environment of a speech sound, is called selective adaptation: an ambiguous stimulus between [ba] and [da] is perceived as [ba] if it is preceded by multiple [da]s, but as [da] if it is preceded by multiple [ba]s. This was explained by the assumption of specialized detectors for features that were subject to fatigue if repeatedly activated by an input; subsequent investigations, however, showed that adaptation can be explained by a combination of simple contrast effects with auditory adaptation. The second such effect, reflecting an influence of the larger-scale environment, is the effect of stimulus set range. If a stimulus set contains a good example of [ba] but only bad examples of [da], listeners will be more likely to identify intermediate stimuli as [da] than if the continuum ranges from a 'good' [da] to a 'bad' [ba]. Recently, a specifically lexically conditioned range effect has been discovered. Norris *et al.* (2003) exposed listeners to two lists of words and nonwords. One list contained some words ending in a 'good' [s] and other words ending with a 'bad' [f], while the other list contained words ending in 'good' [f] and a 'bad' [s]. Similar to the range-effect phenomenon, listeners exposed to the list with the 'good' [s] and the 'bad' [f] were more likely to perceive an ambiguous fricative as [f] than listeners exposed to the other list. In contrast to simple range effects, however, this only occurred if the unclear fricatives were embedded in words and not if they were embedded in nonwords. This indicates that lexical knowledge was used to shift the category boundary. Overall, these effects indicate that listeners have flexible phonetic boundaries, which can change based on influences of the immediate and the overall acoustic–phonetic context as well as based on lexical influences. The fact that languages change over time, and that pronunciation of speech sounds changes not only within a speech community but in the speech of individual members of the community (*see*), further underscores the flexibility of category boundaries.

Context Effects

Listeners' experience of variability in speech produces efficient compensation for variance-adding influences. Ladefoged and Broadbent (1957) presented test words embedded in the sentence *Please say what this word is* to listeners who had to decide whether the test word was *bit*, *bet*, *bat*, or *but*. A test word perceived primarily as *bit* in isolation was perceived predominantly as *bet* when presented in context of a carrier sentence with a lowered first formant. Thus, listeners judged vowels not only by their absolute formant frequencies, but also in relation to the formant frequencies in the surrounding words: if the test word was presented in the sentence with lowered F1, this as it were increased the perceived F1 in the test word, leading to the percept of a lower vowel. (Phonologically low vowels are associated with high F1 frequencies.)

Not only relational cues, such as formant frequencies in the surrounding words, but also intrinsic cues lead to such vowel normalization. For instance, f0 and higher formants may represent a frame against which absolute formant frequencies are evaluated. Other approaches argued for distance or ratio measures (e.g., [F1–f0] or [F2/F1]) as possible invariants over varying absolute values of formant frequencies. In a comparative study, Nearey (1989) showed that the strongest effects seem to be caused by relational properties, although – in keeping with the redundancy and multiple determination of speech perception – intrinsic vowel information also plays a role.

Compensation also occurs for within-speaker variation due to phonetic context. Lindblom and Studdert-Kennedy (1967) found that vowels with the same center frequencies were judged as more [back] – corresponding to a lower perceived F2 – in a [dVd] context than in a [bVb]. Similar effects have been observed for liquid-stop, fricative-stop, fricative-vowel, and vowel-nasal contexts (see Holt and Kluender, 2000 for a review). Likewise, compensatory influences have been observed for speech rate; a formant movement of a given duration is more likely to be perceived as [ba] if it is followed by a long vowel, as [wa] if it followed by a short vowel. A physically long vowel signals a slow rate of speech, so that a transition of a given length appears to signal a relatively shorter formant movement in such a context (Miller and Liberman, 1979).

Although these context effects impressively attest to listeners' ability to undo the introduction of variability in speech production, casual speech still holds sources of variance that challenge models of speech perception, such as segment deletions and reductions of phonological forms. Little is yet known about the perceptual resolution of such processes.

Theories of Speech Perception

The task of any theory of speech perception is to integrate the facts of segment perception as just laid out into a coherent framework. Two questions have been central for theories of speech perception. First, is there an innate specialization for speech perception that sets it apart from 'ordinary' auditory perception? Second, does speech recognition take speech production into account? In a certain way, the answer to the second question has to be yes. Speakers and listeners communicate efficiently, which necessarily implies that the listeners can make sense of the production mechanisms. Theories differ as to whether this parity between speaker and listener also implies that the listener takes into account the mechanism that produces speech in speech perception.

Motor Theory The motor theory of speech perception holds that speech is perceived via recognition of the speaker's intended phonetic gestures. Developed at Haskins Laboratories during the 1950s and 1960s by Liberman and colleagues (see Liberman, 1996: Chap. 1), the theory arose from the team's findings that percepts evoked by speech sounds seemed not tightly coupled to acoustic properties. However, there was correspondence to articulatory gestures; thus uniting the onsets of [ki] and [ku] was the fact that in both cases the tongue body obstructs the airflow in the back oral cavity. Articulatory measurements indicating that speech gestures themselves depend strongly on context resulted in later versions of the theory positing a more abstract invariant, the neural commands that initiated articulation.

According to motor theory, speech perception depends on a specialized module that is part of humans' biological endowment for language, not dissimilar to the echo-location ability that enables bats to navigate in the dark. This module captures all auditory stimuli that seem linguistically relevant before 'ordinary' auditory percepts arise. The module uses an internal speech synthesizer to determine which motor commands are most likely to account for the speech input, a process called analysis by synthesis. Such a mechanism can account for the context dependencies and trading relations described above; the internal synthesizer produces the same dependencies in its resynthesis attempts as occur in natural speech. Due to the emphasis on speech gestures, motor theory also accounts naturally for the influences of visually perceived speech on speech perception (*see* **Audio-visual Speech Processing**).

Additional evidence for motor theory comes from studies that indicate a difference between speech and nonspeech perception. The phenomenon of categorical perception was therefore an important finding for early motor theory. Another effect held to show that 'speech is special' is duplex perception; if the formant transitions of the second and third formants of a CV syllable are presented to one ear and the rest of the syllable to the other, the syllable is perceived as an integrated whole. These formant transitions are heard as nonspeech chirps, just as in isolation; nevertheless, they can be integrated with the rest of the syllable to determine whether the syllable is perceived as, say, [da] vs. [ga]. The signal is effectively heard twice, as speech and as nonspeech (hence the term **du**plex perception). An auditory theory might account for an integration of the signals at both ears, but would seem to predict a triple percept, of each of the two parts separately plus the integrated percept. Duplex perception seems to be better accounted for by a distinction between speech perception and auditory perception.

Motor theory's claim that such phenomena are unique for speech perception and/or human listeners is, however, hard to maintain. With respect to categorical perception, within-category discrimination is often possible, and the sharpness and most probable location of category boundaries seem to be shared between human and nonhuman listeners. Context effects, trading relations, and duplex perception have been observed with nonhuman listeners and/or nonspeech sounds. Such observations do not, of course, contradict the assumption that speech perception rests on specialized mechanisms. The auditory abilities of nonhuman listeners may indeed allow some speech perception, but this does not imply that human listeners perceive speech by the same means. However, positing a very powerful innate speech analyzer is unparsimonious. Further, motor theory suggests that the neural substrate for speech perception differs from that for nonspeech perception. If the capture of a sound by the speech-perception system prevents alternative auditory perception, this module should be located early along the auditory neural pathway. Current investigations, however, indicate a great overlap of speech and nonspeech perception in terms of neural substrate in these early cortical areas (Scott and Wise, 2004).

Direct-Perception Theory The theory of direct perception of speech (e.g., Fowler, 1996) also maintains that the perception of speech is tightly linked to production mechanisms. Based on Gibson's theory of direct perception, the theory holds that perceptual systems have evolved to perceive the sound-causing structures directly using higher-order invariants. A well-known example of an higher-order invariant is a cue for size in vision: an object on a checkered surface covers the same number of squares whatever the distance between observer and object. Similarly, the movements of the articulators are supposed to give rise to higher-order invariants that allow them to be perceived directly. Thus, the direct-perception theory and motor theory both assume that the perceptual system recovers speech gestures from the signal. However, in the direct-perception theory this is achieved without an innately specified speech synthesizer.

By referring to articulation, direct-perception theory accounts for context effects and trading relations in the perception of speech just as motor theory does. According to the direct-perception account, the evidence that motor theorists hold to show a perceptual difference between speech and nonspeech is actually an effect of meaningfulness. The nonspeech sounds under investigation were not environmental sounds, as speech is, but meaningless constructed sounds. The importance of ecological significance of the sounds under investigation was buttressed by the finding that duplex perception occurs for meaningful nonspeech sounds such as a slamming door (see Fowler, 1996). Given that the direct-perception theory does not argue that speech is special, evidence for human-like perception of speech by nonhuman species is not problematic for the theory.

To distinguish direct-perception theory from more general accounts (see below), it is necessary to show that speech-perception phenomena cannot be accounted for by either an auditory or general learning explanation. One piece of such evidence derives from the multimodality of speech perception. The McGurk effect (*see* **Audio-visual Speech Processing**) shows integration of auditory and visual speech, and similar integration arises when perceivers hear and feel – with their hand on the lips of a speaker – a syllable. Just as with visual speech, haptic speech codetermines the ultimate percept. The absence of opportunity to learn the association between felt and heard gestures – we rarely touch each other's lips during conversation – speaks against a general learning account and by implication in favor of the direct-perception account. The principal weakness of the direct-perception account, however, is that it is questionable that higher-order invariants indeed specify speech gestures; little progress has been made in underpinning this part of the theory.

General Auditory and Learning Account The assumption that speech perception relies on general auditory and perceptual-learning abilities (Diehl *et al.*, 2004) is more a framework than a theory. This approach denies specialization for speech perception and excludes reference to production in perception. The challenge for this approach is to account for the exquisite attunement of listeners to the dynamic properties of speech production; it is addressed by reference to evidence of perception-based constraints on production (see, e.g., Ohala, 1996). Despite large differences across languages in the size of vowel inventories, for example, vowel systems seem to place vowels in the vowel space so as to maximize perceptual distance, and hence minimize confusability between vowels. Motor theorists' arguments that the structure of phonetic categories – with a large number of potential cues, each individually unreliable – can only be captured by reference to production is countered by pointing out that categories without invariant features are not unique to speech (consider Wittgenstein's category *game*). Context effects, likewise, are not unique to speech perception, but are typical for perception in general. For some context effects and trading relations, auditory processing seems indeed to be sufficient (*see* **Categorical Perception in Animals**).

From Phonemes to Words

To arrive at an utterance's interpretation, listeners must in effect segment the continuous input and identify the sequence of individual words comprising it. These two processes are interdependent.

Segmentation

Word boundaries in running speech are similar to phonetic categories in that there is no reliable acoustic cue analogous to the white space which marks word boundaries in written text. In Figure 2, there are three visible pauses. All of these pauses occur in front of voiceless stops, and in only one case (the onset of *purposes*) does the pause correspond to a word boundary. Even worse, the word boundary between *this sentence*, does not even coincide with a phoneme boundary, as the two occurrences of /s/, before and after the word boundary, are produced as a single geminate fricative. However, just as phonetic categories make up for the lack of invariance by a large number of potential cues, there are multiple, in large part language-specific, cues to word boundaries.

A first class of cues is physically 'in the signal' and based on fine phonetic detail. For example, voiceless stops in English are generally unaspirated, but in word-initial, position will be aspirated. Figure 2 provides an example with a longer VOT for the first /p/ in *purposes*. Listeners can hence distinguish, for example, *ice cream* from *I scream* based on the phonetic properties of the [k]-sound. Embedded words, such as *cap* in *captain*, may also be recognized on the basis of fine phonetic detail: f0 contour and duration of the syllable [kæp] differ between instances of the intended one-syllable and two-syllable word. This type of effect constrains the possible architecture of prelexical processing: the fine phonetic detail must be retained in order to allow subphonemic detail to influence lexical segmentation (see McQueen *et al.*, 2003).

Currently, there are at least three accounts for how phonetic detail is used. First, an episodic account assumes that there are no prelexical units, but that the lexicon consists of episodic traces of words. For the case of embedded word, a *cap* from *captain* will not produce a good fit with the episodic traces stored for occurrences of the word *cap*. Second, an allophonic model would postulate different phoneme-sized prelexical units for different instances of the same phoneme, depending on position within a syllable. Finally, a prosodic account proposes that fine phonetic detail in f0 and duration are subject to a prosodic analysis, which is used to guide segmentation. Note that the last two accounts are not necessarily competitors, but may capture different, temporal and spectral, aspects of fine phonetic detail.

A second class of cues to word boundaries is statistical in nature. Phonotactic constraints prohibit certain sequences syllable internally (e.g., /mr/ cannot occur syllable internally in English or Dutch). Other sequences are probabilistically associated with word boundaries (e.g., many words or few words begin in such a way). Moreover, such cues need not be restricted to adjacent phonemes: vowel harmony cues in languages such as Finnish can indicate to listeners where word boundaries occur. All such cues are language-specific, as also are the rhythmic cues that listeners use to segment speech (stress rhythm in stress languages, etc.). Listeners are able to exploit all these types of boundary cues (see McQueen and Cutler, 2001).

Finally, lexical competition also contributes to segmentation. Consider the utterance *ship inquiry.* The first two syllables correspond to the word *shipping*, the next to *choir.* This, however, leaves *y* unaccounted for. This parse would be prevented if word candidates compete for activation, and this competition also influences word segmentation. Then, both *ship* and *inquiry* would inhibit the candidate *shipping*, while the parse *shipping* and *quiry* would only contain one lexical candidate that may inhibit activation of other items. All current models of spoken-word recognition incorporate some such process (*see* McQueen and Cutler, 2001, and **Speech Recognition: Psychology Approaches**).

Models of Spoken-Word Recognition

Models of word recognition aim to account for the retrieval of long-term memory information given the perceptual input. There is consensus on at least three issues: activation, parallelism, competition. All models make use of the activation metaphor: word candidates can be activated in a graded fashion. In contrast to look-up from a printed dictionary, the process is not a one-to-one search and comparison between input and stored knowledge, but instead multiple word candidates are activated in parallel. Activated candidates then compete for recognition. The models differ on many issues of architecture, which are discussed in more detail in **Speech Recognition: Psychology Approaches**.

In what follows, we give a short impression of existing models of spoken-word recognition. Each model has its specific merits and weaknesses, as the models differ in their primary motivation. This makes the models partly incommensurable. Models can be compared, however, with respect to their basic units and assumptions.

TRACE TRACE (McClelland and Elman, 1986) assumes that speech input activates phonetic features, which in turn activate phonemes, and these in turn activate word nodes. These three levels are linked by facilitatory connections in both directions. Thus, the feature 'nasal' activates the phoneme nodes [n], [m], and [ng]; the phoneme node [n] activates the word node for *neat*, *nose*, *narrow* etc. Once a word is sufficiently activated, facilitation can flow from the word node back to the phoneme nodes for each segment in the word. Within each level, candidates (features, phonemes, or words) compete with one another via inhibitory connections. TRACE was the first model to posit such an active process of inhibition between word candidates. The most controversial aspect of TRACE is, however, the assumption of facilitating top-down connections from lexicon to prelexical nodes. This form of feedback, however, does not seem to be functional (facilitation flows from the most activated word candidate; as this word is already most activated, it needs no further assistance to win the competition; Norris *et al.*, 2000) and the behavior of the model does not crucially depend on the top-down connections.

In order to account for time invariance, TRACE duplicates the phoneme and word level multiple times. Thus, [n] in the nth 'input slot' activates *nose* in the lexicon aligned with the nth slot, *ant* in the lexicon aligned with the (n–1)th slot, and so on. This solution of the time invariance problem necessitates a forbidding number of connections between word nodes, because every word node is duplicated multiple times and has inhibitory connections to all other word nodes. In consequence, modeling with TRACE is only tractable with a small lexicon.

Shortlist Shortlist (Norris, 1994) models word segmentation and recognition in continuous speech, and allows simulations with a realistic lexicon of tens of thousands of words. These features probably make it the most complete model of word recognition. In its current form, the model's weakness is that it relies, as a simplifying assumption, on a phonemic input pattern.

The model assumes that word units are activated by match with bottom-up input and inhibited by bottom-up mismatch, thus accounting for evidence that even in long words, phonological mismatch in one segment can have powerful effects (Norris *et al.*, 2000). Shortlist differs from TRACE both in this utilization of mismatch information and in the absence of top-down feedback.

Once a number of candidates are activated, a shortlist of candidates is generated, and these are wired into a purpose-built interactive-activation network, where inhibitory connections between candidates induce competition. In order to achieve segmentation, the model uses, besides the competition, attested constraints and perceptually relevant cues to word boundaries (see 'Segmentation').

MINERVA2 Goldinger's (1998) model of an episodic, exemplar-based lexicon assumes no prelexical level; in consequence, the question of feedback does not arise. The episodic lexicon consists of episodes of words in which a 'name' part of a given episode was assumed to be invariant, while other parts encoded varying situational contexts and voices. A given stimulus causes different traces to resonate depending on overlap with the input. If multiple instances of the same word have been presented before, the echo will be relatively ‘abstract’: multiple traces differing in voice and context information will be activated, with voice and context activations canceling each other out. If only a few episodes are available, the echo will entail a large amount of voice and context information. The model accounts well for the amount of phonetic detail that listeners replicate in repeating speech: more phonetic detail is retained for low-frequency words, or nonwords presented only rarely in a training phase, than for high-frequency words, or nonwords presented often in training.

One weakness of the model is that it claims to be a model of word recognition without speaker normalization. In its current form, however, speaker and word information are completely separated in memory. Given that both vowel identity and speaker identity determine, for instance, formant values in vowels, the model implicitly assumes a prelexical level at which vowel identity and speaker identity influences on the speech signal are distinct. This separation is exactly what normalization achieves, and hence the input to the model is not more realistic than the input that is used to drive models involving normalization.

Neighborhood Activation Model The neighborhood activation model (NAM) (Luce and Pisoni, 1998) allows specific predictions of competition between word candidates. In its current form(s), the model is severely limited by being implemented only for words with three segments. It assumes a prelexical processing stage with allophonic recoding of the acoustic input without activating feedback from a lexical level.

The model incorporates a recoding process in which a string of segments as initial input is fed through a known confusion matrix established for CV and VC syllables. This gives rise to graded activations of multiple allophone candidates for every segment, and accordingly, to the activation of multiple candidates. Not only the target word itself is activated, but also its phonetic ‘neighborhood.’ This allows the model to account for evidence from phonetic priming experiments for similarity-based activation and inhibition. The dynamics of the competition process depend on the neighborhood density and the balance of the frequency of the target word and its neighboring competitors. Lexical selection is achieved by a choice rule that takes into account the (frequency-biased) pattern of activation across the competitor set.

Distributed Cohort Model The distributed cohort model (DCM) (Gaskell and Marslen-Wilson, 1997) is currently the only model that relates phonological and semantic representations (though this topic has attracted considerable attention in theories of speech production, *see* **Spoken Language Production: Psycholinguistic Approach**). It posits phonological feature values presented in serial fashion as input, these to be processed by a recurrent network and mapped onto distributed patterns in the lexicon.

Different lexical knowledge domains, namely semantics and phonology, are represented in parallel. Competition between word candidates arises not through lateral inhibition, but because of limited capacity in the lexical networks. Words are not represented by word nodes, but different words make use of the same units in the semantic and phonological networks. It follows that the dynamics will be different in the phonological domain than in the semantic domain. While representations in the phonological domain slowly converge on the phonological representation of a word as it unfolds itself, the semantic network shows a more chaotic behavior. This is a consequence of the fact that words such as *captain* and *captive* share phonological features but have little semantic overlap.

Adaptive Resonance Theory From the framework of adaptive resonance theory (ART), a number of models have been proposed that deal with word recognition and segmentation (e.g., Grossberg and Myers, 2000), though no model with a large vocabulary based on this framework has been presented to date. ART assumes that recognition depends on the development of a resonant cycle between word-level representations and prelexical representations. This involves online feedback from the lexical to the prelexical level activating the corresponding phoneme units, which in turn activate the lexical unit. The nature of this feedback is, however, different from that in TRACE. Here the resonance is crucial because lexical nodes cannot reach more than subliminal levels of activation without it.

The model is able to accept either features or phonemes as input. It uses no ‘slots’ for the input; input is presented in continuous time. In principle, this should allow ART to capture the role of temporal detail in the segmentation and competition process.

Summary

Our understanding of the human perception of speech sounds and recognition of words has grown considerably over the last decades and will continue to grow. It has become apparent that phoneme perception exploits multiple and diverse cues, and draws on immediate and distant acoustic–phonetic context. However, the recognition of speech sounds is probably not achieved without some uncertainty, and as a consequence, the recognition of words entails the activation of multiple candidates that compete for recognition.

At the same time, basic issues have not been resolved. Do we perceive speech as sound or as gestures? Is there a prelexical unit that is used to access the lexicon, and if so, what form(s) should it take? Does the lexicon contain abstract information, or does it consist of a multitude of concrete episodes? The differing theories can be aptly described as research programs that are difficult to compare. Significant progress may be achieved if, for instance, abstractionist models try to account for the evidence that motivates episodic models – exquisite memory for phonetic detail – while episodic models address the evidence that speaks for normalization and for generalization of phonetic knowledge across words. Hybrid models that can account for both these sets of evidence may set the tone for the next generation of word perception models. In the domain of phoneme perception, gestural theories have to address the question of how gestures can actually be perceived, while a more general account needs to accumulate constraints in order to become more testable.

See also: Audio-visual Speech Processing; Categorical Perception in Animals; Speech Production; Speech Recognition: Psychology Approaches; Spoken Language Production: Psycholinguistic Approach.

Bibliography

Diehl R L, Lotto A J & Holt L L (2004). ‘Speech perception.’ *Annual Review of Psychology 55,* 149–179.

Fowler C A (1996). ‘Listeners do hear sounds, not tongues.’ *Journal of the Acoustical Society of America 99,* 1730–1741.

Gaskell G M & Marslen-Wilson W D (1997). ‘Integrating form and meaning: a distributed model of speech perception.’ *Language and Cognitive Processes 12,* 613–656.

Goldinger S D (1998). ‘Echoes of echoes? An episodic theory of lexical access.’ *Psychological Review 105,* 251–279.

Grossberg S & Myers C W (2000). ‘The resonant dynamics of speech perception: interword integration and duration dependent backward effects.’ *Psychological Review 107,* 735–767.

Harnad S (1987). *Categorical perception: the groundwork of cognition.* New York: Cambridge University Press.

Hockett C (1955). *A manual of phonology.* International Journal of American Linguistics Memoir 11. Baltimore: Waverly Press.

Holt L L & Kluender K R (2000). ‘General auditory processes contribute to perceptual accommodation of coarticulation.’ *Phonetica 57,* 170–180.

Klatt D H (1989). ‘Review of selected models of speech perception.’ In Marslen-Wilson W D (ed.) *Lexical representation and process.* Cambridge, MA: MIT Press. 169–226.

Kluender K R, Lotto A J, Holt L L & Bloedel S L (1998). ‘Role of experience for language-specific functional mappings of vowel sounds.’ *Journal of the Acoustical Society of America 104,* 3568–3582.

Kuhl P K (1991). ‘Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not.’ *Perception and Psychophysics 50,* 93–107.

Ladefoged P & Broadbent D E (1957). ‘Information conveyed by vowels.’ *Journal of the Acoustical Society of America 29,* 98–104.

Ladefoged P & Maddieson I (1996). *Sounds of the world’s languages.* Oxford: Blackwell Publishers.

Liberman A M (1996). *Speech: a special code.* Cambridge, MA: MIT Press.

Lindblom B & Studdert-Kennedy M (1967). ‘On the role of formant transitions in vowel recognition.’ *Journal of the Acoustical Society of America 42,* 830–843.

Lisker L (1986). ‘“Voicing” in English: a catalogue of acoustic features signaling /b/ versus /p/ in trochees.’ *Language and Speech 29,* 3–11.

Luce P A & Pisoni D B (1998). ‘Recognizing spoken words: the Neighborhood Activation Model.’ *Ear and Hearing 19,* 1–36.

Maddieson I (1984). *Patterns of sounds.* Cambridge, UK: Cambridge University Press.

McClelland J L & Elman J L (1986). ‘The TRACE model of speech perception.’ *Cognitive Psychology 23,* 1–44.

McQueen J M & Cutler A (2001). ‘Spoken word access processes: an introduction.’ *Language and Cognitive Processes 16,* 469–490.

McQueen J M, Dahan D & Cutler A (2003). ‘Continuity and gradedness in speech processing.’ In Schiller N O & Meyer A S (eds.) *Phonetics and phonology in language comprehension and production: differences and similarities.* Mouton de Gruyter: Berlin. 39–78.

Miller J L & Liberman A M (1979). ‘Some effects of later-occurring information on the perception of stop consonant and semivowel.’ *Perception and Psychophysics 25,* 457–465.

Nearey T D (1989). ‘Static, dynamic, and relational properties in vowel perception.’ *Journal of the Acoustic Society of America 85,* 2088–2113.

Norris D (1994). ‘Shortlist: a connectionistic model of continuous speech recognition.’ *Cognition 52,* 189–234.

Norris D, McQueen J M & Cutler A (2000). ‘Merging information in speech recognition: feedback is never necessary.’ *Behavioral and Brain Sciences 23,* 299–325.

Norris D, McQueen J M & Cutler A (2003). 'Perceptual learning in speech.' *Cognitive Psychology 47,* 204–238.

Ohala J J (1996). 'Speech perception is hearing sound, not tongues.' *Journal of the Acoustic Society of America 99,* 1718–1725.

Scott S K & Wise R J S (2004). 'The functional neuroanatomy of prelexical processing in speech perception.' *Cognition 92,* 13–45.

Strange W (1999). 'Perception of vowels: dynamic constancy.' In Pickett J M (ed.) *The acoustics of speech communication: fundamentals, speech perception theory, and technology.* Boston, MA: Allyn and Bacon.

Sussman H M, Fruchter D, Hilbert J & Sirosh J (1998). 'Linear correlates in the speech signal: the orderly output constraint.' *Behavioral and Brain Sciences 21,* 241–299.

Speech Processes in Dysarthria

F R Boutsen, University of Oklahoma, Oklahoma City, OK, USA

In an era when the terms 'speech' and 'language' were used interchangeably, and Broca and Wernicke's constructs of 'aphemia' and 'aphasia symptom complex' found harmonious reconciliation in localizationist theories of language, Kussmaul (1881) broke tradition by asserting that the seat of speech is most likely not confined to a cerebral convolution. Furthermore, unlike his contemporaries, he drew a clear distinction between the neurological disorders of speech and language. In what is conceivably the first classification of neurogenic communication disorders, he defined as separate from aphasia a group of articulation disorders that were due to organic or psychic disturbances of the central nervous system (CNS). These articulation disorders he labeled dysarthria, which were to be distinguished from the dyslalias that resulted from peripheral lesions and/or malformations of the articulators or the cranial nerves (Grewel, 1957).

Overall, Kussmaul's broad neurological roadmap, albeit provocative for its time, did not much more than delimit the concept of dysarthria, confining it to the CNS apart from language and functional/organic speech disorders. It was not until later that more refined neurological classification schemes began to also assert a coupling between the still fairly amorphous dysarthria symptom complex and etiologies that were bound within levels or components of the central nervous system. For example, Zentay (1937), Froeshels (1943), and Luschsinger and Arnold (1949) classified the etiologies associated with dysarthria at four levels within the central nervous system (the corticobulbar, cortico-strio-pallido-rubro-bulbar, frontopontine, and cerebellar levels). Growing consideration of speech processes, other than 'articulation,' as well as a broadening array of 'dysarthric' symptoms needing theoretical cover, soon stretched and refined the four-level classification schemes to include the peripheral nervous system (PNS) and subdivisions within the neuroanatomic levels. The classifications of Peacher (1950) and Grewel (1957), perhaps, best embodied this trend. In seminal papers, these authors formulated what even by today's standards can be regarded as a modern neurological perspective on dysarthria.

Peacher and Grewel defined dysarthrias, with the exclusion of developmental, somatic, or psychogenic speech disorders, as disturbances of the speaking system resulting from neurological disorders that involve the cortical, subcortical, brainstem, and spinal levels. Physiologically, the aforementioned disorders were proposed to yield distinct motor deficits that in turn patterned different dysarthria types. Those enumerated in Grewel's theory included dysarthrias associated with flaccid or spastic paralysis, rigidity, discoordination, lack of sensation, difficulty with praxis, and disinhibition (as seen in echolalia).

Though conspicuous in some places of their theory, the objective to link the varied phenomenology of dysarthria to neuroanatomical levels was not the sole organizing principle in either of the aforementioned frameworks. This is because the proposed divisional lesions of the nervous system that they argued were capable of framing a dysarthria type were also undergirded by views concerning the neuroanatomical levels of the 'normal' speech mechanism if not the various processes that partake in it. The theoretical requirement that a neuroanatomical map of dysarthria would be consistent with that underlying normal speech processes was perhaps most strongly articulated by Peacher (1950). He argued that while clinically, it might be anticipated that dysarthria presents varying degrees of dysfunction in articulation, phonation, resonation and respiration, speech rhythm needs inclusion as well. He felt its incorporation was justified in view of normal speech data reported by Stetson (1932) who had shown that