Clinical Focus

Hearing and Believing: Some Limits to the Auditory-Perceptual Assessment of Speech and Voice Disorders

Ray D. Kent

Waisman Center on Mental Retardation and Human Development, University of Wisconsin–Madison

Speech-language pathology relies on auditory-perceptual judgment as a central tool for classifying and measuring a variety of disorders of communication. Over the history of the field, a great deal has been written about the use of perceptual judgments for research and clinical practice. Auditory-perceptual methods carry strong advantages of convenience, economy, and robustness, but it is also clear that these judgments are susceptible to a variety of sources of error and bias. Awareness of these threats to validity and reliability is a major step in the effective and refined use of perceptual methods. Several common themes are evident in contemporary research on the perceptual assessment of voice disorders, stuttering, dysarthria, aphasia, and apraxia of speech. These five disorders are taken as primary foci in a discussion that (a) identifies threats to reliability and validity, and (b) offers suggestions for the improvement of auditory-perceptual methods, whether used alone or in combination with instrumental techniques.

The ear is the essential tool of the speech-language pathologist. Auditory-perceptual judgments are typically the final arbiter in clinical decision-making and often provide the standards against which instrumental (so-called “objective”) measures are evaluated. Although it is difficult to know the exact degree of clinicians’ reliance on auditory-perceptual methods, surveys of selected clinical centers indicate that these methods are used frequently and valued highly (Gerratt, Till, Rosenbek, Wertz, & Boysen, 1991). In addition, for many disorders, or for some components of complex communicative disorders, auditory-perceptual judgment is the primary (if not singular) means for evaluating the outcome of an intervention program. As McCauley (1989) observed, measurement of some kind guides decisions at every step of the clinical process, including assessment, intervention, termination of treatment, and follow-up. Because of the critical importance of auditory perception in clinical practice, speech-language clinicians should be keenly aware of the strengths and weaknesses of auditory-perceptual judgments of speech. If shortcomings are identified, it is important to ask:

1. How can these limitations affect clinical decisions?
2. Can these problems be remedied?

The ideal scale should permit reliable discriminations between normal and disordered speech or voice (Dejonckere, Obbens, de Moor, & Wieneke, 1993). It also should permit reliable discriminations of intraspeaker change, such as exacerbation in progressive disorders or improvement of a condition following successful intervention. Clinical applications of auditory judgments are predicated on the assumptions that listeners (a) have a common understanding of perceptual labels such as hoarse, nasal, rough, monoloud, excess and equal stress, or stuttering; (b) use essentially the same verbal descriptors and associated scale values to assess a given sample of speech or voice; (c) can isolate for judgment one perceptual dimension from several co-occurring dimensions; (d) have a uniform reliability in judging the various dimensions that give a complete clinical portrait of speech or voice disorders; and (e) can make perceptual judgments for which the interjudge differences are smaller than the differences needed for clinical classification or to discern changes in clinical status. One should be concerned about auditory-perceptual judgments in clinical practice because the assumptions underlying the use of these judgments do not always characterize reality. Some particular problems are as follows:

1. Judges do not appear to have equivalent definitions of dimensions to be rated.
2. Specialists fail to reach consensus on which perceptual dimensions should be rated for a given disorder.
3. Perceptual ratings of various dimensions are intercorrelated, that is, they are not independent. When this happens, the values obtained for any one dimension may be influenced by co-occurring dimensions of a disorder.

4. Various perceptual dimensions are not rated with uniform reliability.

5. Differences among expert judges are larger than the differences needed for diagnostic classification or the effects of intervention.

This is by no means an exhaustive listing, but simply points to some outcomes that spell concern to those who rely on perceptual assessments. A further reason for including these particular statements is that, as will be discussed, each of them is true in some conditions pertinent to clinical practice.

Sources of Variability in Auditory-Perceptual Judgments

Errors in the perception of speech are part of the normal and ordinary communication experience. It appears that listeners sometimes hear what is not there and sometimes fail to hear what is there. The auditory perception of speech is by no means the only perceptual modality that errs in these ways. Human perceptions—auditory, balance, visual, tactile, olfactory, and gustatory—are vulnerable to a number of failings. Errors in the perception of normal speech are instructive about the tendencies of the ear to “mishear.” Limitations in the perception of normal speech may be exaggerated in the perception of speech that is atypical or disordered.

The word error is used very broadly in this paper to include various types of uncertainty in perceptual processing, as well as illusions or lapses in perception. A fundamental concern is with the grain or resolution of auditory judgments that are typically used in clinical or research settings. A particular limitation is that discriminable differences among stimuli are not necessarily isomorphic with labeling responses, that is, identification judgments. It is frequently the case that listeners can make finer auditory discriminations than they can label using available identification responses. The classic demonstration of this disparity for speech were experiments showing that identification and discrimination functions were dissimilar for simple formant patterns heard as stop + vowel syllables (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). This experimental evidence gave rise to the concept of “categorical perception” of speech sounds. This concept has been revisited several times in the literature, and recent examinations of the problem emphasize psychophysical bases of the phenomenon (Macmillan, 1977; Pastore et al., 1977). Psychophysical factors relevant to speech and voice judgments also are discussed in Handel (1989), Klatt (1982), and Watson (1973). These sources are good background to a discussion of auditory-perceptual judgments of speech and voice.

Auditory Illusions: Phonemic Restoration and Verbal Transformation

One example of a perceptual inaccuracy is demonstrated in Warren’s (1976) work on auditory illusions, especially phonemic restoration and verbal transformation. In the phonemic restoration effect, listeners fail to detect that a speech sound has been replaced by a nonspeech sound such as a cough. This phenomenon is an example of hearing something that really is not there. It presumably reflects the engagement of the listener’s higher-level, top-down (knowledge-based) strategies by which hypotheses are made about semantic and syntactic features of a nascent utterance. The hypotheses are sufficiently important in deriving the meaning of an utterance that acoustic discrepancies may be ignored. In the verbal transformation effect, listeners hear a changing phonetic pattern for an unchanging acoustic stimulus, such as a word that is replayed repeatedly 50–100 times. As the stimulus is replayed, listeners commonly report a changing percept, often hearing an entirely different word. One interpretation of this phenomenon is that the listener is continually advancing hypotheses about an incoming speech signal. The same phenomenon may occur when a clinician repeatedly plays a recorded stretch of speech for the purpose of phonetic transcription or rating of a disorder. The experiments on phonemic restoration and verbal transformation demonstrate that speech perception is at times loosely derived from the acoustic signal. Knowledge-based hypotheses are pervasive influences.

Misperceptions of Natural Fluent Speech

Misperceptions of natural fluent speech have been described by Bond and Garnes (1980). For the most part, these perceptual errors appear to reflect strategies in the perceptual process. Apparently, listeners’ attempts to make linguistic sense of a speech signal carry the seeds for potential errors in perception. The listener’s strategies that support retrieval of a linguistic message include the following:

1. Listen for stress and intonation patterns.
2. Derive a phrase structure.
3. Try to recognize words.
4. Pay special attention to stressed vowels.

These are global or relatively high-level aspects of language comprehension and, as such, they constitute biases in the auditory comprehension of spoken messages. They are useful but also can be sources of error. Bond and Garnes’s analyses of misperceptions revealed that over a quarter of the errors were segmental substitutions, the great majority of which were for single consonants. Place of articulation was particularly likely to be a source of error for consonants. Stressed vowel errors were rare. Some errors were attributable to two different uses of phonological processes: (a) the listener does not realize that a simplification rule has been used by the speaker and therefore fails to recover a reduced segment or syllable, or (b) the listener believes that a rule was used in production and hears a spurious segment or syllable. Other error types involved misplaced word boundaries or lexical confusions.
If these sources of error are robust enough to affect the perception of normal speech, then it is reasonable to expect that they apply as well to disordered speech.

**Phonemic False Evaluation**

Phonemic false evaluation is the mistaken recognition of phonemes that were not in fact produced by the talker (Buckingham & Yule, 1987). Listeners hear speech through a number of sieves, each of which accomplishes a type of information filtering. Of particular interest are those situations in which listeners hear a sound for which there is no basis in the acoustic signal. Buckingham and Yule (1987) reviewed evidence for phonemic false evaluation in linguistic transcriptions of different languages. The gist of their discussion is that phonetic-level differences can lead to phonemic differences in a transcription of language behavior. Obviously, errors of this kind hold serious implications for phonological studies of developing and disordered speech.

Normal speech is not error-free speech, and several reports have been published on corpora of speech errors such as spoonerisms. The detection of errors in normal speech is an interesting test case for auditory-phonetic perception, given that the errors presumably appear as isolated events in the context of otherwise normal speech. A number of strong claims have been made about the nature of these errors and their implications for theories of speech regulation at the phonological, phonetic, and motor levels (Fromkin, 1971; MacKay, 1970; Shattuck-Hufnagel, 1983; Stemberger, 1982). For example, it frequently has been concluded that the errors arise largely or entirely at levels above that of motor commands. These analyses are important in their implications for the detection of errors in atypical or disordered speech as well, all the more so because similar methods of error analysis often are used. Mowrey and MacKay (1990) questioned the validity and reliability of virtually the entire literature on normal speech errors. Their analyses pointed to three primary limitations in the traditional perceptual classification of speech errors: (a) listener normalization is a regular part of the perceptual process that may override detection of subtle errors; (b) errors at a fine phonetic level are difficult to detect reliably, especially when the listener is attending to verbal content; and (c) suitable transcription techniques are lacking for highly or subtly anomalous sounds. Mowrey and MacKay concluded that, “the problem of error characterization is so pervasive and its effect so great as to render the significance of traditionally collected data corpora questionable” (p. 1311). There seems to be no reason why the same objection should not be raised with respect to the corpora of errors reported in the clinical literature on aphasia, apraxia of speech, and children’s misarticulations. This point will be taken up again later.

**Effects of Lexical Status on Phonetic Categorization**

Effects of lexical status on phonetic categorization have been reported in a number of studies (Burton, Baum, & Blumstein, 1989; Fox, 1984; Ganong, 1980). Briefly, the lexical status of a sound stimulus can affect the phonetic boundary for a constituent feature such as voice onset time (VOT). For example, when VOT boundaries are compared for the word *dash* and the nonword *tash*, the former commands a relatively greater range of VOT values. One might say that the acoustic-phonetic space is carved to serve lexical decision-making. The general term for this effect is the lexical identification shift. The implication of this effect for clinical studies is that the clinician’s lexical biases may influence the ability to detect allophonic differences in sound production. In a similar phenomenon, the semantic content of a sentence can influence the perception of acoustically ambiguous words (Gaines & Bond, 1976). It is likely that some words in disordered speech are acoustically ambiguous, and that semantic context can play a deciding role in clinical judgments of lexical acceptability.

**Equivalence Classes in Phonetic Perception**

Generally, an equivalence class is a category of related sounds. The members of the category differ somewhat in their physical attributes, but they all are sufficient to evoke the category in question. Therefore, class members may be discriminable under certain conditions, but they all are associated with a common identification response. The phoneme is frequently given as an example of an equivalence class in speech.

Equivalence classes in the perception of sounds in native (L1) and non-native (L2) languages have been the focus of a number of perceptual experiments. One conceptualization of equivalence classes is from Flege (1987). Flege states that when two sounds from different languages are acoustically similar, the sound from the non-native language tends to be assimilated to (treated as equivalent to) the sound in the native language. But when two sounds from different languages are dissimilar acoustically, equivalence is not as likely to occur. Flege (in press) has developed a Speech Learning Model to explain how phonetic elements of a non-native language are acquired. This model emphasizes perceptual representation as the key factor in learning the L2 phonological system.

Best (1993, 1994) proposed a perceptual assimilation model for the discrimination of non-native phonetic contrasts. A basic tenet of the model is that listeners perceptually assimilate non-native contrasts to the phonemic categories of their native language. The direction and degree of assimilation varies with the similarity between the native and non-native phones in question. Assimilation can take the form of five basic patterns, which are described below for the example of two non-native phones to be discriminated:

**Two-category contrasts** are those in which each non-native phone is assimilated to a different native phoneme category. For these contrasts, excellent discrimination is expected because each phone is attached to a familiar category. **Single-category contrasts** are contrasts in which both non-native phones are equally assimilable to the same native phoneme category. Poor discrimination is predicted
for this situation, given that the listener is essentially making a within-category discrimination. Uncategorizable contrasts involve sounds that are heard as speech but neither phone can be assimilated to a native phoneme category. Discrimination is predicted to be poor but better than for the single-category contrasts. Category goodness difference contrasts refer to contrasts in which both non-native phones are assimilated to the same phoneme category but the two items differ in their goodness of fit to that category. Discrimination is expected to be moderate to good. Nonassimilated contrasts involve non-native phones, both of which are outside of the native phonetic space and may not even be heard as speech. Discrimination is expected to be moderate to good.

The perceptual assimilation model was developed to account for variations in the discrimination of non-native phonetic contrasts and is therefore relevant to second-language learning. However, the basic ideas would seem to apply as well to situations in which a clinician judges the speech of a client with an atypical development of the native language.

McGurk Effect: Seeing and/or Listening

In the McGurk effect, visual and auditory information can interact in phonetic decision-making (MacDonald & McGurk, 1978; McGurk & MacDonald, 1976). For example, it has been demonstrated that a signal that is unambiguously identified as /da/ in an auditory-only presentation may be heard as /ba/ when accompanied by simultaneous video information showing labial closure. Another example is the perception of /d/ as the result of a visually presented /b/ and an auditorily presented /g/. The listener apparently synthesizes the visual and auditory cues to derive a phonetic decision. Visual information may override or complement auditory information, even if the latter is sufficient for reliable phonemic identification in the absence of visual cues. This phenomenon is evidence that speech perception makes use of available visual information.

Prosodic Influences on Phonetic Classification

Several types of prosodic variation might influence phonetic decisions, but one in particular that has been described in experimental reports is speaking rate. As speaking rate is changed, the acoustic boundaries of phonetic segments can change as well. Wayland and Miller (1994) concluded that slowing of sentential speaking rate shifts the locations of best exemplars of VOT ranges. As the authors state, “sentence-level rate information…alters the internal perceptual structure of a phonetic category” (p. 2700). This point is of considerable interest to the study of speech disorders such as dysarthria that are typically associated with a slowed speaking rate. If speaking rate and other prosodic variables influence phonetic decisions in normal speech, then it is highly likely that similar prosodic influences are at work in the perception of disordered speech, as Weismer and Martin (1992) discuss with respect to the dysarthrias.

Talker-Listener-Utterance Interactions

A series of studies summarized by Bradlow, Nygaard, and Pisoni (1995) point to three general conclusions. First, variations in intelligibility are observed for different sentences across many talkers and listeners. Second, variations in intelligibility are observed for different talkers across many sentences and listeners. Third, variations are observed in the perceptual strategies that different listeners apply in learning to identify different talkers, and in the way this talker-specific information is used in speech perception. Talkers, listeners, and utterances interact in determining decisions about intelligibility and talker identification. Especially to the point of the present paper, different normal talkers are not equally intelligible, and listeners differ in the strategies they use in listening to different talkers. The study of these variations has been rather neglected, but they are the focus of several recent papers reviewed by Bradlow and colleagues (1995).

Conclusion

Errors or variations in the perception of normal speech take a number of forms. These probably occur with some frequency in everyday experience and point to limitations in the accuracy of perceptual processing. But it should be emphasized that not all of these limitations are undesirable. Some of them reflect the cognitive processes of spoken language perception and account, for example, for a listener’s ability to understand speech in noisy situations. If reliability is defined as the stability of a measurement or detection process, then auditory-perceptual processing is not perfectly reliable and may be highly transient in some analyses that it performs on the speech signal.

Experiments on speech perception have demonstrated that segmental-phonetic decisions can be influenced by accompanying prosodic information and that listeners are sensitive to various types of prosodic and paralinguistic information (Mullenix & Pisoni, 1990; Nusbaum & Morin, 1992). These results are consistent with a parallel-contingent mode of speech perception, in which acoustic information is extracted from the speech signal to supply two information paths, one for segmental-phonetic information and the other for suprasegmental (prosodic and paralinguistic) information. Although the paths for phonetic and suprasegmental information are parallel, phonetic decisions are ultimately based on contingent information from the suprasegmental path. That is, phonetic decisions can be influenced by suprasegmental information such as speaking rate and speaker identity. In this view, interaction of segmental and suprasegmental information is inevitable.

Limits on the Sensitivity and Reliability of Auditory Assessments of Speech and Voice Disorders

No attempt will be made here to review the immense literature that has accumulated on the auditory-perceptual assessment of speech and voice. But reference will be
made to contemporary papers that exemplify the issues. The following comments pertain to selected problem areas in the perceptual assessment of disordered speech or voice. The areas include: disordered voice, stuttering, dysarthria, aphasia, and apraxia of speech. First, however, brief comments will be made on the problem of rating normal voices. A consideration of this area raises issues that are directly pertinent to the assessment of voice disorders.

**Judging Normal Voices**

One of the most basic questions to ask is: What should the clinician scale or rate in perceptual tasks—what dimensions, qualities, or features? Further, it should be asked if these are the same for normal and disordered speakers. Although it might be taken for granted that normal and disordered speech and voice can be assessed using a common system, there is reason to question this assumption.

Gelper’s (1988) research is an example of attempts to determine the appropriate perceptual rating scales for normal voices. She concluded that some attributes were scaled more reliably than others. Both listener confidence in assigning ratings and actual interjudge agreement were high for the dimensions of pitch, loudness, and rate of speech. Interjudge agreement also was relatively high for vocal variability and age. Gelper noted that perceptual rating systems designed for use with disordered speech or voice may not be effective in distinguishing among normal variants of speech and voice. A related question is: What is a normal voice? Is there really one normal standard that might be defined by the intersection of several dimensions? Or are there really varieties of normal vocal behavior rather than a single normal template?

Multidimensional scaling analysis is one way to identify the number and nature of dimensions used in perceptual judgments of voices. Gelfer (1993) used this method to study the ratings of 20 female voices by two groups of listeners, a group of speech-language pathologists and a group of untrained listeners. The data for the speech-language pathologists were satisfactorily analyzed with a 5-dimensional solution (the dimensions were perceived pitch, loudness, age and rate, variability in pitch, and voice quality). The solution for the data from the untrained listeners had only two dimensions. These results indicate that experience may help the listener to judge reliably a greater number of properties in the voice. Unfortunately, not much is known about how experience affects the use of the various sources of information in the complex acoustic signal of speech or voice. Quality is a complex multidimensional attribute, and much remains to be done to define its acoustic correlates and to model the ways in which experience affects quality judgments.

**Judging Disordered Voices**

The situation may be even more confused in the assessment of disordered speech. Perkins (1971) identified 27 different terms that were used in nine studies of abnormal voice quality. Of the total 27 terms, only two (hoarse and nasal) were used in all nine studies. This lack of convergence on basic descriptors is a fundamental obstacle in the standardization of voice ratings. Various solutions to the problem of voice rating have been described. Voice rating systems for clinical use typically use from 4 to 20 basic dimensions. However, Laver’s (1980) componential analysis uses about 40 basic “settings” to describe a voice quality. Each setting can be further described with a scalar label. There is considerable variation among systems with respect to the dimensions that are used to judge voice quality.

One obstacle in voice ratings is that judges may disagree among themselves on which aspects of voice quality are most important in rating normal and disordered voices. Kreiman, Gerratt, and Precoda (1990) concluded that, “clinical training and experience cause listeners to differ more, not less, in how they perceive voice quality, at least in tasks that involve unstructured similarity judgments” (p. 109). This conclusion runs counter to the frequent assumption that training will consistently yield improved agreement among judges. Kreiman and colleagues also noted that, “The differences between clinicians were large enough to suggest that averaging data across subjects may produce misleading results and obscure important aspects of an individual subject’s perceptual behavior” (p. 109). In a later paper, Kreiman, Gerratt, and Berke (1994) specifically investigated ratings of “breathy” and “rough” qualities in disordered voices. They concluded that the two qualities are related multidimensional constructs and that judgments of roughness depended strongly on concomitant breathiness but not vice versa. Similarly, Wolfe and Ratusnik (1988) reported that judgments of vocal pitch interacted with roughness. Vowels that were considered to be moderately to severely dysphonic were matched with a significantly lower pitch than were vowels that were considered clear to mildly dysphonic. For example, rough vowels were matched with a pitch that was on the average six semitones lower than the fundamental frequency. Pitch judgments are influenced by several factors. Although vocal fundamental frequency is a major correlate in many studies, the strength of the correlation between pitch judgments and vocal fundamental frequency may well vary with different voice qualities.

**Judging Stuttering**

A central issue in stuttering research is the measurement of the disorder, and a huge literature has accumulated on this problem. Suffice it to say that this issue has a continued prominence in published papers and conference discussions (see, for example, Perkins, 1990, and commentaries). In his review, Young (1984) concluded, “Agreement among observers for identifying stuttering instances, or for consistently differentiating stuttering instances from normal disfluencies, is very low. The same pattern of poor agreement holds for individual observers on repeated trials” (p. 28). Moreover, Young noted that agreement was not improved when observers were given behavioral definitions of stuttering, when both visual and auditory information was available to them, or when the rate of
speech was slowed electronically. Similarly, Cordes and Ingham (1994c) concluded that, “observers disagree markedly when they are asked to identify stuttering” and “the measurement systems most commonly used in stuttering treatment and research cannot be depended upon to produce reliable data about stuttering” (p. 1295). An additional difficulty is that the procedures most frequently used to estimate reliability may not be suitable for the task. Lewis (1994) surveyed articles on stuttering published in three national journals over a 5-year period. The author identified 55 articles for which observer agreement was regarded to be important. Lewis noted that 93% of the surveyed articles either did not report an index of observer agreement or relied on one of five indices that were judged to be seriously flawed. Lewis added that, “Of the studies in which observer agreement values were reported, 92% relied on one of five misapplied indices and nearly all those dealt with establishing observer agreement on nominal category judgments” (p. 277).

The likelihood of a judgment of stuttering for a speech sample varies with the constituent disfluency types. The disfluency types most likely to be labeled as stuttering are sound, syllable, and word-level repetitions and, to a lesser degree, prolongations (Young, 1984). In a recent study, Althaus, Goorhuis-Brouwer, Kloosterman, Vink, and Minderera (1993) reported that different types of disfluency had different values of reliability in perceptual determination. Reliable perception was noted for elongations, blockages, repetitions, and interjections; unreliable perception was reported for length of pauses and quality of breathing during phonation.

Cordes and Ingham (1994a, 1994b, 1994c) concluded that the inconsistency in counting stuttering events is larger than some reported treatment effects. The inconsistency does not appear to disappear when experienced judges are used. Ingham and Cordes (1992) presented data showing both interclinic differences and interjudge differences in the identification of stuttering events. Cordes and Ingham (1995) reported that judgments of stuttering by “10 highly experienced authorities on stuttering treatment and research” (p. 33) showed considerable variation across sites. These studies underscore the limitations of auditory-perceptual assessments in judging the adequacy of the clinician’s tools for assessing a disorder or for measuring change in clinical intervention.

**Judging Dysarthrias**

The perceptual rating scales central to the Mayo Clinic system for assessing the dysarthrias are perhaps the most comprehensive set of rating dimensions in speech-language pathology (Darley, Aronson, & Brown, 1969a, 1969b, 1975a). These dimensions address a number of qualities related to voice, articulation, resonance, and prosody. The comprehensiveness of these scales gives them a heuristic value to communication disorders generally. It is of special interest, then, to examine recent experience with the perceptual assessment of the dysarthrias.

Since the appearance of the Mayo Clinic system in 1969, there have been very few studies of its validity and reliability. But there is evidence that the attributes in this system are not scaled with equal reliability and that judges may differ in the selection of salient perceptual features to describe a given disorder. Zyski and Weisiger (1987) used recorded samples of dysarthria from materials prepared by Darley, Aronson, and Brown (1975b) to determine the accuracy with which classification of type of dysarthria was identified by three groups of listeners who varied in experience. Listener Group 1 is of particular relevance: The group consisted of 17 speech-language pathologists who had a minimum of 5 years of clinical experience and who routinely diagnosed and treated persons with dysarthria. The 17 listeners identified type of dysarthria for 28 recorded samples. Their accuracy was 19%, ranging from a low of 1% for flaccid dysarthria to 55% for hypokinetic dysarthria. Another group of judges were graduate students who received 5 hours of classroom training on the perceptual evaluation of dysarthria. The graduate students did better than the professionals. They had an accuracy of 56%, or a little better than half of the samples correctly identified. Perhaps their superior performance reflected the fact that they were trained with the Audio Seminar materials prepared by Darley, Aronson, and Brown. Zyski and Weisiger (1987) reached the sobering conclusion that the Mayo Clinic system was not sufficiently reliable for clinical purposes.

Zeplin and Kent (1996) also replicated the rating studies of Darley and colleagues using the recorded materials prepared by Darley and colleagues for a training tape. Zeplin and Kent reported only partial agreement with the most deviant perceptual dimensions listed in the original study by Darley and colleagues for six forms of dysarthria. Although Zeplin and Kent’s data on the reliability of their five judges generally indicated fairly high intrajudge and interjudge consistency, the reliability was not as high as might be desired for all dimensions. Among those dimensions that frequently appeared as the ten most deviant across types of dysarthria (that is, dimensions with particular importance in characterizing dysarthrias as a disorder group), consistently small standard deviations were observed for **imprecise consonants**, **pitch level**, **loudness**, and **fast rate**. In contrast, **monopitch** and **monoloudness** had relatively large standard deviations for some dysarthrias.

The optimum dimensions for auditory-perceptual rating may vary among speaking tasks. Zeplin and Kent (1996) observed differences between the tasks of syllable repetition and passage reading. When the ten most deviant dimensions were compared for the two tasks, the number of shared dimensions varied from five to seven across types of dysarthria. Of particular interest are the dysarthrias for which the top five ranked dimensions were the same for the contextual and syllable-repetition tasks. Only the two types of hyperkinetic dysarthria met this criterion.

Sheard, Adams, and Davis (1991) examined the performance of 15 speech-language pathologists in rating the speech of 15 persons with ataxic dysarthria on five dimensions: **imprecise consonants**, **excess and equal stress**, **irregular articulatory breakdown**, **distorted vowels**, and **harsh voice**. The agreement between judges was lowest for...
the dimensions of irregular articulatory breakdown and harsh voice. Sheard and colleagues reported generally high intercorrelations between the rated dimensions and interpreted this result to mean that the dimensions are not perceptually independent. The dimension of harsh voice was the least correlated with other dimensions. The authors concluded that, “clinicians should be aware that the actual rating values they assign to apparently separate dimensions may in fact reflect an overall perception of a number of concurrent, crucial/salient speech characteristics” (p. 291).

Related issues in the use of perceptual dimensions were discussed by Southwood and Weismer (1993) and Kim (1994). Southwood and Weismer collected perceptual ratings for the four dimensions of acceptability, naturalness, normalcy, and bizarreness for the speech of 18 male talkers with amyotrophic lateral sclerosis (ALS). The perceptual ratings for the four dimensions were strongly interrelated, and Southwood and Weismer suggested two interpretations of this result: (a) the listeners may not have perceived these nominal dimensions as distinct qualities of dysarthric speech in ALS, or (b) the scaling of all four dimensions depended on severity to such an extent that independent judgments could not be obtained with the methods used. Because intelligibility was used as an overall index of severity, the authors drew a conclusion with important clinical implications: intelligibility is a good indicator of overall speech impairment (although intelligibility is not necessarily a good indicator of voice impairment). And, because intelligibility correlates so highly with dimensions such as acceptability or naturalness, intelligibility may be a premium dimension for an efficient perceptual protocol.

Kim (1994) examined the acoustic and perceptual bases of monotonyness in speech production in individuals with Parkinson’s disease. She concluded that listeners’ ratings could not distinguish among the three monotonous dimensions of monopitch, monoloudness, and monoduration. Furthermore, monopitch and monoloudness were strongly correlated (r = 0.98). Kim’s results raise questions about the degree to which the perceptual dimensions of monopitch and monoloudness, as used in the Mayo Clinic rating system, can be independently rated by listeners.

It has been suggested that listening to the speech of individuals with dysarthria may be similar to listening to degraded speech signals in the sense that in both of these conditions, the listener relies on higher-level cognitive processing to retrieve the intended message (Tjaden & Liss, 1995b). Some studies have shown that the intelligibility of dysarthric speech varies with the predictability of its syntactic and semantic content (Hammen, Yorkston, & Dowden, 1991) and with listeners’ familiarity with the dysarthric patterns (Platt, Andrews, Young, & Quinn, 1980; Tjaden & Liss, 1995b).

Transcribing Errors in Aphasia and Apraxia of Speech

Buckingham and Yule (1987) wrote, “The aphasic is not a speaker of another language obviously, but hearers are not usually equipped to catch certain subtle articulatory aberrations brought about by brain damage just as they are not equipped to catch certain articulations brought about by virtue of the fact that some speaker is speaking an unfamiliar language to the hearer, or, as in the Japanese case, where the speaker is substituting sounds from his native language for English sounds when speaking English” (p. 117). This quotation sets the stage for a consideration of perceptual biases and other errors as they relate to clinical practice. The literature on phonemic and phonetic analyses of aphasia and apraxia of speech is too extensive to be reviewed with any satisfaction here, but it is useful to take a quick retrospective look at aphasia and apraxia of speech, the understanding of which have evolved with the methods used to study them.

Phonemic paraphasias are frequently recorded in the speech of individuals with aphasia. In one of the most careful investigations of this phenomenon, Shinn and Blumstein (1983) identified phonemic paraphasias through unanimous agreement of four phonetically trained judges. Their criterion pruned the potential number of phonemic paraphasias from 240 to 9 in Broca aphasia and from 60 to 2 in Wernicke aphasia. In their summary of this and related studies, Buckingham and Yule (1987) wrote, “there is likely to be a subtle articulatory disruption in the fluent aphasics that affects the acoustic picture, which leads to hearer inconsistency and to false evaluation of phonemic intentions of the fluent aphasic speakers” (p. 120).

Transcriptions of apraxia of speech originally provided evidence of a predominance of substitution errors, and interpretations of the disorder accordingly emphasized phonemic-level mechanisms. But data from acoustic and physiologic studies indicated that apraxia of speech is characterized by frequent subtle errors—errors that apparently missed perceptual attention. (For a review of this literature, see McNeil & Kent, 1990; Rosenbek, Kent, & LaPointe, 1984; Rosenbek & McNeil, 1991). As these errors were documented and confirmed, a different view of apraxia of speech emerged. The lesson here is that it can be very easy to confuse level of description with level of explanation. Just because phonemic errors are easy to detect does not mean that the disorder process necessarily operates exclusively or primarily at a phonemic level of speech organization. As mentioned earlier, essentially the same objection has been raised against virtually all of the published corpora on errors in normal speech (Mowrey & MacKay, 1990).

Some General Procedural Issues

Mode of Stimulus Presentation

The field of psychophysics is concerned in large part with the effects of methodology on perception, such as the way in which the delivery of stimuli influences the judgments of the stimuli. Several reports indicate that the reliability of perceptual judgments of speech can be affected by the method of stimulus presentation. Stephens and Daniloff (1977) reported lower reliability values and a greater range of scores under audio-recorded versus live conditions. McNutt, Wicki, and Paulsen (1991) observed a larger number of judged phoneme errors and a greater
variance of judgments in auditory-only versus auditory-visual presentation of the stimuli. The availability of visual information and the quality of the acoustic signal can be important factors in perceptual judgments of speech.

**Differences in Auditory Judgment for Various Sound Classes**

Although firm conclusions are not easily drawn, it appears that judgments of misarticulation are not equally reliable for all sound classes. Judgments appear to be more reliable for /s/ than /r/ (Elbert, Shelton, & Arndt, 1967; Hoffman & Schuckers, 1978; McNutt et al., 1991; Stephens & Daniloff, 1977). However, this difference was not confirmed by Shriberg and Lof (1991).

**Narrow Versus Broad Perceptual Judgments**

Another area of concern is the questionable use of complex coding systems to derive detailed information from perceptual judgments. The complexity typically takes the form of increasing the number of response categories or making fine distinctions within individual response categories. The goal of such added complexity is to increase information from perceptual analysis, but the supplementary information may be of questionable reliability. The reliability of phonetic transcription has been reported to be poor for narrow as opposed to broad transcription (Pye, Wilcox, & Siren, 1988; Shriberg & Lof, 1991). The conclusion seems to be: Narrow transcription is important because it promises increased information about the behavior (speech articulation) to be described, but it also seems to be accompanied by poor reliability and therefore a form of diminished information transfer (greater uncertainty). For additional discussion of issues of reliability in phonetic transcription, see Shriberg and Kent (1995).

The finding that narrow phonetic transcription is less reliable than broad transcription parallels a result reported by Shaw and Coggins (1991) who evaluated the reliability of Ling’s (1976) phonetic-level evaluation for children with hearing loss. Shaw and Coggins concluded from their use of this procedure with a group of children with severe and profound hearing loss that it may not have acceptable levels of interobserver agreement. They remarked that the complexity of the coding system appeared to reduce the reliability of the phonetic level evaluation.

Building detail and complexity into a coding system does not guarantee that the listener’s auditory-perceptual decision space is equal to the demands of the analysis task. The most efficient and reliable coding system is one that is closely matched to the capacity of auditory discriminations and auditory memory. It is well to recall George Miller’s (1956) famous “magic number” of seven plus or minus two, which is a commonly encountered limit on the number of perceptual decisions that can be reliably made along a given perceptual continuum. The limits to human information processing are real and the use of complex and subtle decision categories can exceed the practical limits of auditory discriminations.

A related problem is that coding systems developed for typical adult speech may have a lower reliability and uncertain validity when used with developing or disordered speech. Infant vocalizations are one example of extending the application of one coding system, the International Phonetic Alphabet, beyond its traditional domain. Stockman, Woods, and Tishman (1981) studied the reliability of three individuals with “extensive teaching and field experiences in phonetics” (p. 597) to transcribe 1,032 vocalizations obtained from four female babies within the age range of 7–21 months. Stockman and colleagues concluded that when the transcriptions were compared with an identical segment match criterion, both interjudge and intrajudge agreement were generally below 60% of the total number of segment comparisons made at any age within the 7–21 month range. Agreement tended to increase with increases in the age at which the samples were recorded, but the overall low level of reliability should be taken into account in evaluating phonetic transcription data for infant vocalizations. This limitation is relevant to the general study of early speech given that data of this type constitute the bulk of the research literature on infant vocalizations (Kent & Miolo, 1994).

**Differences Among Rating Scales**

The choice of rating scale can affect the judge’s ability to discern reliable differences along a dimension of interest. To take one example relevant to a series of studies in the literature on speech, a given dimension may be conceived as prothetic or metathetic (Stevens, 1975). A prothetic dimension is one that varies in magnitude or quantity. It is sometimes described as an additive, quantitative continuum. In this kind of dimension, the perceptual judgment is one of determining if more or less of an attribute is present. Loudness is often given as an example of a prothetic scale. A metathetic scale is one that varies in terms of a change in quality. It is sometimes described as a substitutive, qualitative continuum. Pitch is given by some writers as an example of a metathetic dimension. A metathetic dimension can be scaled with either an equal-appearing interval scale (EAIS) or a direct-magnitude estimation scale (DMES), but a prothetic dimension is scaled more satisfactorily by the DMES. A number of studies have demonstrated how these two kinds of perceptual continua relate to dimensions important to the assessment of speech and voice (Metz, Schiavetti, & Sacco, 1990; Samar & Metz, 1988; Schiavetti, 1984; Schiavetti, 1992; Schiavetti, Metz, & Sitler, 1981). It is likely that some multidimensional rating instruments, such as the system of Darley and colleagues (1969a) for rating dysarthric speech, combine prothetic and metathetic scales. This issue will be taken up later in this paper with a consideration of alternative procedures suitable for the perceptual assessment of multidimensional signals.

**Biasing Effects of Speaker Characteristics**

Auditory-perceptual assessment may be biased by certain speaker characteristics, such as physical appearance.
and history (Lopez, 1989; Ramig, 1975; Scheuerle, Guilford, & Garcia, 1982). Even trained clinicians can be unwittingly influenced by these factors. The implication is that perceptual judgments, or interpretations of these judgments, by a clinician can be influenced by a client’s physical features and by knowledge of a client’s history.

Auditory Salience of Speech Characteristics

A study by Seikel, Wilcox, and Davis (1990) indicated that clinicians’ judgments of severity may depend primarily on the slowly varying components of the temporal pattern of speech and rather less on the rapidly varying temporal features such as stop bursts and voice onset times. In addition, a study by Zeplin and Kent (1996) showed that the perceptual dimensions judged most deviant in dysarthric speech can vary as a function of the speech material. Auditory-perceptual judgments are based on manifold factors that are poorly understood, but it is likely that some factors are more influential than others. De Krom (1994) reported that ratings of voice quality were more reliable for vowel-onset and whole-vowel stimuli than for post-onset stimuli, apparently because cues for voice quality are less salient in the most stable portion of vowel phonation.

Listener Characteristics

Many listener characteristics are potentially important, but one particular example is the listener’s familiarity with the speaker. Several writers have commented on the influence of familiarity on the listener’s ability to understand a speaker with a speech disorder or to identify errors in that person’s speech (Brodkey, 1972; Doyle, Swift, & Haaf, 1989; McGarr, 1983; Platt et al., 1980; Tjaden & Liss, 1995a, 1995b; Weist & Kruppe, 1977; Wilcox, Kouri, & Caswell, 1990). In general, the evidence is that listeners can adapt to atypical patterns of speech production.

Another listener characteristic is linguistic experience. All listeners come to the task of speech assessment with certain linguistic backgrounds that can powerfully affect perceptual judgments. Some of the clearest evidence comes from studies of phonological disorders in children. Several studies have shown that these children occasionally produce acoustically distinctive productions of sounds that were not perceptually differentiated by examiners (Catts & Jensen, 1983; Chaney, 1988; Forrest & Rockman, 1988; Gierut & Dinnse, 1986; Maxwell & Weism, 1982; Powell, Elbert, & Forrest, 1990; Smit & Bernthal, 1983; Tyler, Edwards, & Saxman, 1990; Weism, Dinnse, & Elbert, 1981). The lack of perceptual differentiation of the sounds can lead the clinician to conclude that the child has a collapsed phonetic contrast. But, in fact, the child is producing a contrast that is not detected by the adult examiner.

Influence of Higher-Level Information on Phonetic Judgments

A good example of the influence of phonetic expectation on phonetic transcription was an experiment reported by Oller and Eilers (1975), in which six judges trained in phonetic transcription transcribed utterances from a 4-year-old boy with delayed phonological development. A comparison was made between transcriptions made with and without knowledge of the meaning of the utterances to be transcribed. The results clearly demonstrated an effect of phonetic expectation based on meaning of the utterances. When the judges knew the meanings of the utterances, their transcriptions conformed more closely to the expected adult forms of the utterances. Given the phonemic restoration effect discussed earlier, it would not be surprising if some phonetic corrections were completely overlooked when the utterance contains highly deterministic syntactic and semantic information.

Conclusion

A number of factors limit the validity and reliability of perceptual judgments of speech and voice disorders. The foregoing list is certainly not exhaustive, but it suffices to show that the problems are manifold and often serious. Moreover, they would seem to apply in some degree to virtually any speech or voice disorder. The depth of the problems cannot be fully explored in this paper; for related discussions of reliability and validity in clinical judgments, see Bassich and Ludlow (1986), Cordes (1994), Kearns and Simmons (1988), Shriberg (1972), and Tosi and Bertoccini (1990).

Procedures to Improve the Sensitivity and Reliability of Auditory-Perceptual Assessment

A first step is to recognize the sources of variability in auditory-perceptual judgments. A host of factors can influence the rating of speech or voice samples. This is not to say that auditory-perceptual judgments are dispensable. For many purposes in speech-language pathology, there is simply nothing better, at least not at this time. Moreover, auditory-perceptual assessment has a long history of successful application to various issues in the clinic and the research laboratory. Among the advantages of the auditory-perceptual method are convenience, economy, usefulness for outcome assessment, and robustness. The first three are obvious, but the third deserves further comment. A remarkable feature of the auditory system is its ability to understand speech under various conditions of interference (masking, temporal interruptions, filtering) and to assess disorders of speech and voice over a range of severity. Instrumental methods do not always compare with auditory judgments in respect to robustness. Rabinov, Kreiman, Gerratt, and Bielamowicz (1995) observed that acoustic measures of jitter can be unusable with severe voice disorders. Some algorithms for extraction of fundamental frequency (or fundamental period in the time domain) have little resistance to noise and therefore will not work satisfactorily for all clients seen by a speech-language clinician. Acoustic analysis can be highly informative, but it also can be misleading if the signals to be analyzed are not suitable for the analysis to be performed. Furthermore,
it is becoming clear that voices are not equal with respect
to the derivaton of reliable or representative values.
Scherer, Vail, and Guo (1995) reported that the number of
tokens required to derive representative voice perturbation
values varies with the stability of the voice sample.
The long experience with perceptual judgments also has
produced ample evidence that these judgments are subject
to various types of error and variability. Unfortunately, the
use of perceptual judgments is not always taught with an
appropriate emphasis on their limitations or the selection of
alternative methods if certain assumptions are not satisfied.
For example, it is often assumed that the various dimen-
sions used to assess voice disorders, stuttering, and
dysarthria are independent and equally reliable. Neither
seems to be true. In fact, it may be prudent to assume that,
in general, perceptual judgments performed on multiple
dimensions will be influenced by strong intercorrelations
among dimensions and by substantial variations in reliabil-
ity across dimensions. Some particular limitations and
suggestions for improvement follow.

Distinguishing of Level of Analysis
From Level of Interpretation

Any analysis has a degree of resolution. A rough-
gained analysis of speech and voice disorders should be
used with care to avoid inappropriate interpretations of the
nature of the disorder. For example, phonemic-level
descriptions can be quite handily used in transcribing the
errors in many neurological disorders of communication
(or even the occasional errors in normal speech), but the
data may not truly reflect the various levels of impairment.
Auditory perception may classify errors into phonemic
categories when a closer inspection could reveal subtle
phonetic (allophonic) differences among the items placed
in a single category. Strong biases operate to reduce the
listener’s sensitivity to subphonemic errors or irregularities

Reference Samples

Kreiman, Gerratt, Kempster, Erman, and Berke (1993)
put the problem of perceptual judgment for voice disorders
in bald terms: “even highly experienced listeners fre-
ly disagree completely about what they hear” (p. 33).
They proposed that “variability in voice quality ratings
might be reduced by replacing listeners’ idiosyncratic,
unstable, internal standards with fixed external standards
for ‘reference voices’ for different vocal qualities” (p. 33).
Although Kreiman and colleagues addressed these com-
ments to the problem of clinical assessment of voice
disorders, they appear to be applicable to a wide range of
communicative disorders. It cannot be taken for granted
that any given term used in perceptual assessment will
have the same meaning for any two judges, or that two
judges will share a verbal description of a clinical speech
sample. Interjudge agreement seems to be better for end-
scale judgments (either normal or highly different from
normal) than for midscale judgments. One means of
selecting reference samples for clinical training is to pick
those that are associated with high levels of interjudge
agreement (Cordes & Ingham, 1995).
Dysarthria is of particular interest because it is one of the
few speech disorders with an arguably dominant classifica-
tion and rating system, the Mayo Clinic system of Darley
and colleagues (1969a, 1969b, 1975a). In addition, generations of
speech-language clinicians were exposed to audiotaped
samples that could constitute standard references for classifi-
cation. The tape by Darley and colleagues (1975b) and its
successor (Aronson, 1993) give us “reference voices” for
selected classifications of dysarthria. Because speech-
language clinicians assess speech disorders primarily, and
often exclusively, by perceptual methods, audiotaped and
videotaped references are needed to standardize the rating
and classification of the dysarthrias, as well as to provide
uniform practice samples. The reference standards used in
conjunction with careful specification of the perceptual
dimensions to be rated should promote listener agreement.
The superior performance of the graduate-student group in
Zyski and Weisiger’s (1987) study may well reflect the
impact of careful training on exemplars. If so, then an
important inference to be drawn from the Zyski and Weisiger
experiment is that training with reference samples may
promote better interjudge agreement than simply selecting
experienced judges.

Procedures of Measurement

The issue of measurement is central to virtually all of
the issues raised in this paper. Although no single approach
to measurement will resolve the many problems that
plague perceptual judgment, there are several procedures
that may generate refined information on perceptual
characteristics of speech and voice. It is beyond the scope
of this paper to consider these approaches in detail, but
some of them can be described briefly as they relate to
selected clinical problems, beginning with stuttering.
A vast proportion of the literature on stuttering pertains to
a central and formidable problem: How should it be mea-
sured? Several authorities believe that the measurement issue
remains a serious obstacle, especially because of the low
levels of interjudge agreement in identifying disfluencies.
Cordes and Ingham (1994a, 1994b, 1994c) identified two
procedural features associated with improved levels of
agreement: (a) measuring stuttering as events in time, rather
than as events tied to individual words or syllables, and (b)
using observers who are highly experienced or trained to
some criterion as part of the study.
A common weakness of many perceptual ratings is that
they are not suited to the multidimensional nature of speech
and voice. Measurement techniques have been developed
that provide a means of analyzing the multidimensionality of
these signals and determining the perceptual independence
(or nonindependence) of the dimensions that constitute the
overall quality impression. Some of this work is based on the
trace-context theory (Braida & Durlach, 1988; Durlach &
Braida, 1969; Macmillan, 1987; Macmillan, Goldberg, &
Braida, 1988). This theory is based on principles of signal
detection theory and Thurstonian scaling. A primary tenet of
trace-context theory is that two types of noise determine the
resolution of judgments on a decision axis. The first type of noise is sensory noise, which is associated with the inherent neural noise of a sensory channel. The second type of noise is memory noise, which varies with the memory load experienced by the subject in different experimental paradigms. This memory noise has two forms. The first form is exemplified in single-interval tasks such as absolute identification. In this type of task, listeners’ judgments are made in a “context-coding” mode such that a stimulus is remembered in a rather rough form as a verbal label defined relative to the overall stimulus context. Perceptual anchors, such as highly familiar categories or stimuli extrema on a continuum, strongly determine the amount of this form of noise. For example, the closer the stimulus to be judged is to a pole of the continuum, the less the noise. The second form of noise is relevant to pairwise discrimination tasks, for which listeners try to hold onto an image (“trace”) of the first stimulus to permit comparison with the second. This form of noise varies with the interstimulus interval. Noise increases as the interval lengthens.

The trace-context theory has been evaluated in experiments on vowel and consonant perception (Macmillan et al., 1988), absolute and relative pitch perception (Burns & Campbell, 1994), and multiple cues of rise time, spectral slope, and cutoff frequency (Campbell, 1994). This line of research has the potential of delineating the dimensions that constitute a multidimensional signal such as speech.

Other work on complex sound perception with direct relevance to speech is based on the expectation that perception is organized with respect to a source-filter model, which is the central assumption of the acoustic theory of speech production (Fant, 1960). Li and Pastore (1995) confirmed this hypothesis by showing that listeners performed more poorly in spectral slope discrimination when fundamental frequency was varied. Because both spectral slope and fundamental frequency are source properties, their covariation hindered listeners’ judgments. But listeners did not show poorer discrimination of spectral slope when spectral shape (a filter property) was varied. The implication is that discrimination of acoustic properties for naturally occurring sounds is determined in part by a source-filter model of perception. An intriguing aspect of this idea is that perception and production are linked by a common organization. This model also carries the implication that discrimination of any given voice (source) characteristic will suffer if other voice characteristics vary. Because vocal pathologies frequently involve simultaneous changes in two or more source characteristics, it would be expected that listeners’ judgments of voice disorders would reflect this limitation on discrimination for any given source characteristic. In addition, it would be expected that simultaneously varying filter (vocal tract) characteristics, as are likely to occur in many individuals with dysarthria, would present difficulties for the discrimination of any single filter characteristic.

Coding Complexity

More is sometimes less when perceptual judgments are concerned. Listeners are not always able to use multivalued or multidimensional scales with adequate reliability. Subtle perceptual discriminations or categorizations can be associated with unacceptably low reliability both in interjudge and intrajudge comparisons. In addition, it seems prudent to assume the nonindependence of many, if not most, perceptual dimensions. The high degree of interdependence among perceptual judgments is justification for the simplification of many complex, multidimensional systems. Research on the interdependence of perceptual judgments should be helpful in pruning the dimensions to a maximally informative set. Multiple dimensions not only require more time from the clinician, but they may yield diminishing returns if the various dimensions are highly correlated.

Influence of Biasing Factors

Because a variety of factors can influence auditory-perceptual judgments, care should be taken to specify the conditions in which such judgments are made. Brief summaries of the listening environment and speech sample should be included as a routine part of clinical reports intended for documentation of disorder or treatment outcome.

Qualifications and Certification of Judges

The clinical research literature frequently establishes the competence of judges employed in perceptual tasks with phrases such as “experienced clinicians,” “persons with substantial experience in listening to speech disorders,” or “certified speech-language pathologists.” Research reviewed in this paper raises serious questions about the degree to which experience alone can be assumed to guarantee satisfactory levels of interjudge agreement. Counter-evidence to this assumption has been published for ratings of voice disorders (Gerratt, Kreiman, Antonanzas-Barrosa, & Berke, 1993; Kreiman et al., 1990; Kreiman, Gerratt, Precoda, & Berke, 1992; Kreiman et al., 1993), stuttering (Cordes & Ingham, 1995), and dysarthria (Zyski & Weisiger, 1987). Experience is valuable for many reasons, but it does not seem to be prima facie evidence of interjudge agreement. Experience seems to lead to interjudge agreement if the experience ensures fundamental commonalities. It would not be surprising if Darley and colleagues (1969a, 1969b) had a better interjudge agreement than would three speech-language pathologists selected from three different clinical sites. Darley, Aronson, and Brown undoubtedly had a substantial amount of shared perceptual experience in developing their 38-dimensional analysis. Their working relationship would be a natural avenue for mutual calibration.

Improving the Correspondence Between Perceptual and Instrumental Analyses of Speech

One way of overcoming the limitations of the perceptual analysis of speech or voice is to supplement perceptual ratings with instrumental (acoustic or physiologic) analyses.
of the same behaviors. Presumably, the instrumental methods would offer a greater reliability and perhaps improved precision. But the potential value of instrumental procedures is still being proved. A brief examination of selected acoustic methods reveals the obstacles to an efficient combination of perceptual and instrumental methods.

The general problem in correlating perceptual and acoustic analyses has been that many studies have shown only weak associations between acoustic measures and perceptual ratings. Discouraging results have been reported for the speech of the deaf (Arends, Povel, Van Os, & Speth, 1990), voice disorders (Eskenazi, Childers, & Hicks, 1990; Wolfe, Fitch, & Cornell, 1995), and dysarthria (Kent et al., 1994; Zwirner, Murry, & Woodson, 1991). Wolfe and Steinfatt (1987) reported that different acoustic variables were the best predictors for different voice types. Kreiman and colleagues (1992) observed that expert judges differed considerably with respect to the acoustic factors correlated with their perceptual ratings of pathological voices. These results indicate that a fixed set of acoustic measures will not necessarily correlate highly with perceived severity across a range of vocal abnormalities and across different judges (even experienced ones). This situation would severely compromise the traditional clinical evaluations, which typically involve a small number of judges (often one) evaluating several different types of speech and voice qualities.

Some steps that have potential to improve the correlations between instrumental and perceptual measures are described below.

**Psychoacoustic Factors in the Choice of Scales of Analysis**

A simple but potentially important step in improving the correlations between perceptual and acoustic data is to use appropriate frequency and intensity scales in the acoustic analysis (Hermansky, 1990; Hermes & Van Gestel, 1991; Syrdal & Gopal, 1986). One alternative is to use acoustic analyses that incorporate psychoacoustic characteristics determined by research on human audition. The Bark scale has attracted considerable attention in basic psychoacoustics and speech sciences. It is now available as an alternative to linear frequency scales in some speech analysis systems developed for personal computers (Read, Buder, & Kent, 1992). Alternatives to the Bark scale also have been described, and some of these may prove preferable for certain purposes. The equivalent-rectangular-bandwidth-rate (ERB-rate) scale may hold advantages for the analysis of intonation, apparently because the critical bandwidths of the ERB-rate scale are narrower than the Bark scale for low frequencies. The ERB-rate scale is similar to the Bark scale for higher frequencies (including most formant frequencies) but is superior for low frequencies and hence for analysis of intonation and for the F1 of high vowels. The ERB-rate scale might serve as a psychophysically based scale that is suited to both the low-frequency energy of intonation and the higher-frequency information of vowel formants and consonant noise energy. Another technique that helps to make both fundamental-frequency and formant-frequency information available on the same display is the wave collation visual speech display (Mitchell & Easton, 1995).

**Improving the Congruence of Perceptual and Acoustic Indices**

Some disparities between perceptual and acoustic indices occur for perceptual judgments that are multidimensional in their acoustic correlates. Kreiman and colleagues (1992) observed that expert judges differed considerably with respect to the acoustic factors correlated with their perceptual ratings of pathological voices. These results indicate that a fixed set of acoustic measures will not necessarily correlate highly with perceived severity across a range of vocal abnormalities and across different judges (even experienced ones). It may be necessary to recognize patterns of perceptual classification even for individual voices. Self-organizing systems are one way to approach this problem (Leinonen, Kangas, Torkkola, & Juvas, 1992). Generally, these systems create ordered patterns from data that may appear on superficial examination to be chaotic or disordered. Self-organizing systems may be a particularly useful way to understand individual differences in perception, such as the differences among expert judges, each of whom may have developed a stable, but distinctive, system for perceptual analysis.

Disparities between perceptual and acoustic data can occur for other reasons. As reviewed earlier, a number of studies have shown that phonologically disordered children sometimes will produce sounds that can be differentiated by acoustic analysis, but not by perceptual judgment. For example, a child may produce voiced and voiceless stops that have a consistent difference in VOT, measured acoustically, but that are judged by the examiner to sound identical. The perceptual judgments lead to the conclusion that the child possesses the contrast but does not realize it in accord with the adult model. VOT is a good example of an acoustic measure that can be used to validate perception. A number of phonetic contrasts are associated with one or two primary acoustic correlates. An inventory of these perceptual-acoustic correlates would be a helpful guide in establishing perceptual-acoustic validation. To say that research on acoustic-perceptual correlates is difficult is not to say that it is impossible or unproductive.

One example of new approaches to acoustic analysis of speech is the use of statistical analyses of the power density spectrum (Buder, Kent, Kent, Milenkovic, & Workinger, 1996; Forrest, Weismer, Hodge, Dinnsen, & Elbert, 1990; Forrest, Weismer, Milenkovic, & Dougall, 1988). The work by Forrest and colleagues showed that the coefficients for three spectral moments (mean, skewness, and kurtosis) could be used to classify noise spectra for stops and fricatives in both normal and disordered speech. Buder and colleagues (in press) explored the use of four spectral moments (mean, standard deviation, skewness,
and kurtosis) in a multiparameter acoustic analysis applied to samples of connected speech. Other acoustic parameters in this analysis are the formant frequencies of the first three formants (derived by automatic LPC formant tracking), fundamental frequency, and rms amplitude. Analyses of this kind could be a useful supplement to perceptual judgments of both segmental and suprasegmental properties of speech. Buder and colleagues described applications to dysarthric speech and demonstrated that the analyses are sensitive to changes in speech that occur with disease progression (specifically, amyotrophic lateral sclerosis) and with clinical treatment (palatal lift prosthesis). Because acoustic analysis can be tedious, especially when multiple measures must be obtained from spectrograms, automatic analysis would hold great advantages in providing fast and reliable data with a minimum of human effort.

A more thoroughgoing application of acoustic analysis to speech disorders has been discouraged by several factors, including the unsuitability of some conventional analyses for the speech of women and children (Kent & Read, 1992), a limited normative database for women and children, and uncertainties in inferring articulatory information from acoustic data. But recent improvements in acoustic analysis systems (Read, Buder, & Kent, 1992), a gradually growing normative database for a wide range of speakers, and improved interpretations of acoustic data should lead to increasingly successful applications of acoustic analysis. As an example of issues in interpretation, Hewlett (1988) demonstrated how acoustic analyses of children’s speech could separate features such as coarticulation, segment duration, and variability in the production of acoustic features such as spectral shape.

It may be helpful to conceptualize the acoustic-perceptual relation in different ways, two of which are polar extremes. At one pole is the many-to-one relation, in which several acoustic variables are associated with a single perceptual attribute, such as hoarseness. A given judgment of degree of hoarseness by a given listener can be associated with a combination of acoustic factors. At the other pole is a one-to-one (or few-to-one) relation, in which a perceptual decision can be mapped against a single acoustic dimension, or against a weighted average of two or more acoustic dimensions. Consonant voicing contrasts are examples of this relation, because the perception of the syllable-initial voicing contrast often can be related to the single acoustic variable of VOT, and the perception of syllable-final voicing contrast frequently can be related to the duration of the vowel that precedes the consonant.

As noted earlier, judgments of severity may depend primarily on the slowly varying components of the temporal pattern of speech and rather less on the rapidly varying temporal features such as stop bursts and voice onset times (Seikel, Wilcox, & Davis, 1990). If such a result has general application, then acoustic analysis of slowly varying temporal properties, such as vowel and consonant durations, could be used to define physical correlates of perceived severity. Acoustic analysis of rapidly varying properties could provide information on the dysarthria that is not as readily appreciated by auditory-perceptual methods.

Expert perceptual skill in humans may be characterized by intuitive rather than analytic performance (Dreyfus & Dreyfus, 1986). As Sheard and colleagues (1991) explain, “in many skilled situations, a person’s efficient analysis of complex stimuli does not seem to be performed according to a set of decision rules on which a computer program might be designed to operate with comparable success” (p. 292). Alternatively, it may be that the current understanding of perceptual decision rules is simply inadequate to permit an adequate computer modeling of the process by which humans reach perception judgments.

**General Conclusion**

Investigators of several speech disorders appear to be converging on some principles to improve auditory-perceptual evaluation. These principles are emerging in research on voice disorders, stuttering, and dysarthria, among other communication disorders. They deserve careful consideration as they pertain to the clinical use of auditory-perceptual methods and to research that employs auditory-perceptual techniques. Some steps for improvement will require deliberate action if they are to be successful on a national scale. For example, reference samples of voice disorders, stuttering, and dysarthria require careful selection and quantification followed by planned exposure, preferably under conditions that encourage proficiency to be established over a number of training sessions.

Auditory-perceptual assessment is critical to much of the clinical practice of speech-language pathology. But the universal application of this kind of assessment is accompanied by certain failings and limitations. Knowing the nature of these failings and limitations is important in minimizing their undesirable effects. Virtually any application of perceptual assessment for any communication disorder runs some risk of false or unreliable information.

Steps to correct for some shortcomings of auditory-perceptual judgment are becoming increasingly defined in the literature, and some common directions for improvement are emerging. Certainly, much remains to be done, and the neglect of basic issues is unfortunate. As Kreiman, Gerratt, and Berke (1994) observed, “The issues of what qualities are perceptually real and perceptually independent have been too long ignored in a field that is founded largely on perception” (p. 1301).

This paper has concentrated on auditory-perceptual analysis because of its long and frequent application in speech-language pathology. However, many of the same concerns apply to visual-perceptual analyses as they pertain to the use of tools such as videofluoroscopy, nasendoscopy, and stroboscopy. Techniques for the visual examination of oral, pharyngeal, velopharyngeal, and laryngeal function usually are associated with a variety of perceptual decisions. Auditory- and visuo-perceptual decisions frequently are the basis for clinical action in speech-language pathology, and, as McCauley (1989) cautioned, “the quality of clinical action depends upon the quality of measurement” (p. 29).
Acknowledgments

This research was supported in part by NIH research grant DC00319 ("Intelligibility Studies of Dysarthria") from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health.

References


Maxwell, E., & Weismer, G. (1982). The contribution of
phonological, acoustic, and perceptual techniques to the characterization of a misarticulating child’s voice contrast for stops. Applied Psycholinguistics, 3, 29–43.


Received March 13, 1995
Accepted February 5, 1996

Contact author: Ray D. Kent, Waisman Center on Mental Retardation and Human Development, University of Wisconsin-Madison, 1500 Highland Avenue, Madison, WI 53705-2280
Email: kentray@waisman.wisc.edu

**Key Words:** speech disorders, voice disorders, perception, auditory-perceptual judgments