In M. Traxler and M. A. Gernsbacher (Eds.), *The Handbook of Psycholinguistics, 2nd Edition* (pp. 201-248). New York: Academic Press, 2006.

Chapter 7

The Perception of Speech

Jennifer S. Pardo and Robert E. Remez

A talker can expect a listener to grasp the rough dimension of any sincere and appropriate message, though only by saying it. For talker and listener, speech is a medium, a link in a commonplace causal chain by which pleasantries or philosophies are exchanged, cooperation is negotiated and compliance is compelled. But, does an essay about speech belong in a book about language? To a newcomer, it is self-evident that conversational partners know what each other says simply by hearing the sounds of spoken words. From this perspective, the fundamentals of speech perception surely lie in psychoacoustics, an essential reduction of speech perception to sensory resolution and auditory categorization. Even so, the newcomer might already notice the difference in auditory quality in the speech of children and adults, or in face to face and in telephone speech, and suspect that the perception of spoken messages entails more than acute hearing. To the old hand familiar with cognitive psychology and the historic place of speech within it, the motivation to study speech perception might seem well and truly relieved now that affordable devices transcribe words from sound. On the contrary, this essay like its companions in this volume was produced by a typing hand and not by a dictating voice, despite the mathematical ingenuity of the engineers – far exceeding that of cognitive psychologists – who create speech-to-text devices. For the reader of any degree of experience, this part of the *Handbook* explains why the descriptive and theoretical puzzles provoked by speech perception have proven to be so enduring, psychologically and linguistically, and in doing so claims a role for speech in language.

Our characterization of the perception of speech ranges across three of its facets. First, we discuss the historic aim of research on speech, which has been to understand how acoustic properties evoke an impression of linguistic form. This line of research is mature, and a sizeable literature beginning with classical sources presents a consistent expression of competing views and evidence. Ideological commitments aside, it is a singular merit of this research tradition that it introduced a generous assortment of theoretical conceptualizations to perceptual psychology. Even when innovation happened to spring from other sources, the well established techniques and research paradigms within the study of speech perception provided a ready means to calibrate the explanatory

Handbook of Psycholinguistics: 2nd Edition
ISBN: 0-12-369374-8

adequacy of a principle. This portion of the essay exposes contemporary viewpoints about perceptual organization and analysis of speech and notes the questions that lead the research forward.

Second, the ordinary perceptual resolution of the linguistic properties of speech is accompanied by an irreducible impression of the talker as well as the message. Research about the recognition of individuals from their speech takes its origin in forensic projects – studies to determine whether a known talker and an unidentified talker are the same – and in artifactual methods to create a vocal identification technology. In contrast to these humble roots, more recent cognitive studies emphasize the perceptual effects of variation in phonetic form across individuals and instances. The evident perceptual interchange of linguistic, individual (or, indexical) and situated properties promised to overturn the classic conceptualization of the acoustic-to-phonetic projection, and this portion of our essay describes the partial success of this project and the questions that remain for a complete causal account.

The third section of our essay characterizes self-regulatory speech perception in which an individual talker's self-perception modulates the production of speech. This theme is contrary to Lashley's founding arguments in psycholinguistics. He held that the rate of production of vocal actions was too rapid to permit monitoring by proprioception, and many studies since have recounted adequate unmonitored articulation, for instance, concurrent to mandibular somatic sensory blockade. This literature about the control of coordination in vocal movement is supplemented and elaborated by more recent studies that identified effects of self-monitoring in other sensory modalities. These findings show that talkers adjust subtle – and, less subtle – properties of articulatory expression as a consequence of phonetic perception, albeit at a slower pace than Lashley stipulated, and in varied social conditions.

Throughout, our essay is organized by psycholinguistic questions, rather than by concerns with specific research methods. Although the investigations that we describe are largely the yield of functional studies of normal adults, we have referred to research about special populations or using special methods when we aimed to secure premises in our argument. We also direct the reader to other discussions when technical matters or special perceivers hold intrinsic interest or importance.

1. PERCEPTUAL ORGANIZATION AND ANALYSIS OF SPEECH

A listener intent on grasping a talker's message must sample physical effects of speech that vary regularly if unpredictably, a consequence of a talker's vocal acts. The regularity as well as the unpredictability derive from a common cause; the linguistic governance of speech deploys formal attributes designated in the talker's language, and these drive the regularities. At the same time, no expression is an exact repetition of a prior one, and whether the departures from stereotypy are attributed to chance or to a specific cause – to a talker's enthusiasm, or haste, or influenza – exact patterns never recur. The central

problem in research on speech has been to understand how perception of regular linguistic attributes is evoked by such unpredictably varying acoustic causes.

None of the acoustic constituents of speech is unique to speech, although some features of speech are characteristic: a cyclical rise and fall of energy associated with a train of syllables, amplitude peaks and valleys in the short-term spectrum, and variation over time in the frequency at which the peaks and valleys occur (Stevens & Blumstein, 1981). In addition to noting these attributes, it is fair to say that natural speech is an acoustic composite of whistles, clicks, hisses, buzzes and hums, a discontinuous and often aperiodic result of the continuous movement of articulators. In following a speech signal, a listener tracks an intermittent pattern of heterogeneous acoustic constituents; there is no single element nor set of them that defines speech, therefore, no simple way for a perceiver to distinguish speech piecemeal from the acoustic effects of other sources of sound. Despite all, a perceiver often tracks the speech of a specific talker sampling by ear and eye, two kinds of *perceptual organization* that also combine multimodally, and resolves the linguistic properties in the sensory effects – that is to say, *perceptual analysis* of the symbolic properties of speech succeeds. We discuss these in turn.

1.1. Perceptual Organization

The ability to track an individual's speech amid other sounds retains the characterization applied long ago by Cherry (1953), the *cocktail party problem*. Such get-togethers can pose many challenges for participants; this specific cocktail party problem is solved by perceivers who understand spoken messages despite the concurrent intrusions of acoustic elements very much like those composing the target speech stream. The sources of unrelated sounds surely include the clinking of glasses and popping of corks, although other extraneous acoustic moments are similar to an attended speech stream because they come from the speech of other talkers. Indoors, the direct sound mixes with late arriving reflections from the ceiling, floor and walls of the attended speech signal itself.

To gauge the means of resolving the sound produced by a single individual, the contrast between visual and auditory attention is instructive. In attending to a visible object or event, a perceiver typically turns to face it, bringing the light reflected by the object of interest to the fovea of the retina. In this retinal region, receptors are densest and pattern acuity is best, for which reasons visual attention will often coincide with a foveated object. A listener's attention to the audible world achieves spatial and spectral focus psychologically, without the selective benefit of a heading at which auditory pattern acuity peaks. In addition, the visible world contains opaque, translucent and transparent objects; the audible world is largely transparent. A listener cannot presume that a sound arriving from a certain direction stems from the visible object at the same heading, for sounds produced by other sources at the same direction are likely to propagate around intermediate objects to impinge on the listener. Despite all, perception often reciprocates the patterned variation of a speech stream with its discontinuities, dissimilarities among components and similarities between its components and those of unattended utterances and other

events. This perceptual function is fast, unlearned, keyed to complex patterns of sensory variation, tolerant of anomalous sensory quality, nonsymbolic and dependent on attention whether elicited or exerted (Remez, Rubin, Berns, Pardo, & Lang, 1994). The evidence to characterize the function and the limits of its effectiveness stems from several lines of research.

1.1.1. Fast

Whether speech occurs in the clear or in noise, it is quickly resolved perceptually if it is resolved at all. Classic studies of the persistence of the auditory trace of speech indicate such fast resolution, for they show that discrimination based on an auditory form of speech becomes poor very rapidly. Before the sensory trace fades, the auditory effects of speech are resolved into a coherent perceptual stream. The estimates of the rate of decay vary, though we can be certain that little of the raw auditory impression of speech is available after 100 ms (Elliott, 1962); and, none after 400 ms (Howell & Darwin, 1977; Pisoni & Tash, 1974). For the perceiver, the perishable auditory form creates an urgent limit on integration of the diverse constituents of speech; auditory properties available to perception are simply lost if integration is delayed. For a theorist, the evident long-term adaptive flexibility exhibited in natural perception cannot be attributed to unelaborated representations of the auditory features of speech without denying this basic psychoacoustic limit (see Grossberg, 2003). In contrast to the natural perceiver, urgency does not constrain artefactual recognizers. The schemes that they employ inherently surpass the physiological characteristics of an auditory system. They can sample and hold acoustic representations of speech analogous to the initial auditory sensory forms; indeed, they can hold them as long as electricity powers the memory (Klatt, 1989). Such superhuman systems have had wide theoretical influence despite indifference to the critical first step of urgent perceptual organization (Picheny, 2003).

1.1.2. Unlearned

Evidence that perceptual organization of speech is unlearned derives from studies of 14 week old infants, who integrated acoustic elements of speech composed through synthesis to be both spectrally and spatially disparate (Eimas & Miller, 1992; cf. Hollich, Newman, & Jusczyk, 2005). Listeners at this young age are hardly aware of linguistic properties in the speech they apprehend, and the perceptual coherence of the diverse constituents can be attributed to precocious sensitivity to vocalization independent of phonetic impressions, and well in advance of linguistic sensitivity. If experience plays a bootstrapping role in perceptual organization during the first three months of life, this is unlikely to entail arduous tutelage, nor sleep learning via exposure to adults whispering in the nursery.

1.1.3. Keyed to complex patterns of sensory variation

The amplitude peaks and valleys in the spectrum of speech are natural resonances of the column of air enclosed within the anatomy of the upper airway. These resonances, or

formants, are set ringing by the regular pulsing of the larynx, which produces harmonic excitation; or, by the production and release of air pressure behind an approximation or occlusion, as in the case of stop consonants; or, by sustained turbulence, as in the case of frication and aspiration. Acoustic changes in the spectrum are nonuniform across the formants. Specifically, the independent control of the articulators that produces formant frequency variation causes uncorrelated differences across the formants in the extent, rise and fall and temporal relation of frequency and amplitude change. Equal change in the first, second, third, nasal and fricative formants is uncharacteristic of vocal sound production, and aggregation of the sensory correlates of speech in perceptual organization occurs without evident reliance on similarity of change across the resonances. In some acoustic transforms of speech spectra, the frequency variation of the resonances is obscured without loss of perceptual coherence. In one version aiming to model the diminished frequency resolution imposed by an electrocochlear prosthesis (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995) the coarse shape of the short-term spectrum envelope was represented in the power of 3 or 4 noise bands, each spanning a large portion of the frequencies of speech. Over time, asynchronous amplitude variation across the noise bands creates a derivative of speech without harmonic excitation and broadband formants, yet the frequency contours of individual resonances are absent. The effectiveness of such variants of speech spectra exposes the basis for the perceptual organization of speech, which lies in detecting a sensory pattern that coincides with phonologically governed articulation. Although these remain to be characterized formally, to a first approximation it is clear that the patterns of sensory variation are complex.

1.1.4. Tolerance of anomalous auditory quality

Perceptual organization of speech is tolerant of anomalous auditory quality, as research with sinewave replicas of utterances has shown (Remez, Rubin, Pisoni, & Carrell, 1981). In this synthetic acoustic signal, the natural products of vocalization are eliminated by imposing the pattern of a speech spectrum on elements that are not vocal in origin. Precisely, three or four pure tones are set to vary in frequency and amplitude in the pattern of the estimated formant peaks of a speech sample. A fourth tone is intermittently used to replicate fricative formants, brief bursts or nasal murmurs. The integration of the tones to compose an intelligible utterance occurs despite the persistence of the weird quality of a sinewave voice, evidence that neither natural acoustic correlates of speech nor auditory impressions of a legitimate voice are required for perceptual organization to occur. Studies with chimerical signals provide independent corroboration that the perceptual organization of speech is indifferent to the specific acoustic constituents of a signal and to the nonvocal auditory quality that can result (Smith, Delgutte, & Oxenham, 2002). To create an acoustic chimera, a coarse grain representation of the spectrum envelope of speech is excited with an arbitrarily chosen source. The result is a composite exhibiting the influence of each aspect, the spoken utterance and the arbitrary source. Like tone analogs of speech, a chimera is intelligible linguistically, evidence that its constituents are grouped to compose a signal fit to analyze as speech. Phenomenally, it retains the quality of the excitation, whether noisy, or harmonic, or, indeed, multiple, as in the instance shown in Figure 1, for which the excitation was provided from an acoustic

sample of a musical ensemble. In each of these critical cases, the perceptual coherence survived the inventory of arbitrary and nonvocal short-term properties by tracking the time-varying properties, which derived from speech even when the acoustic elements did not. This series of findings eliminates as implausible any characterization of perception warranting meticulous attention to elemental speech sounds or their correlated qualitative effects. Instead, they make evident the causal properties of a spectrotemporal pattern superordinate to the momentary constituents.

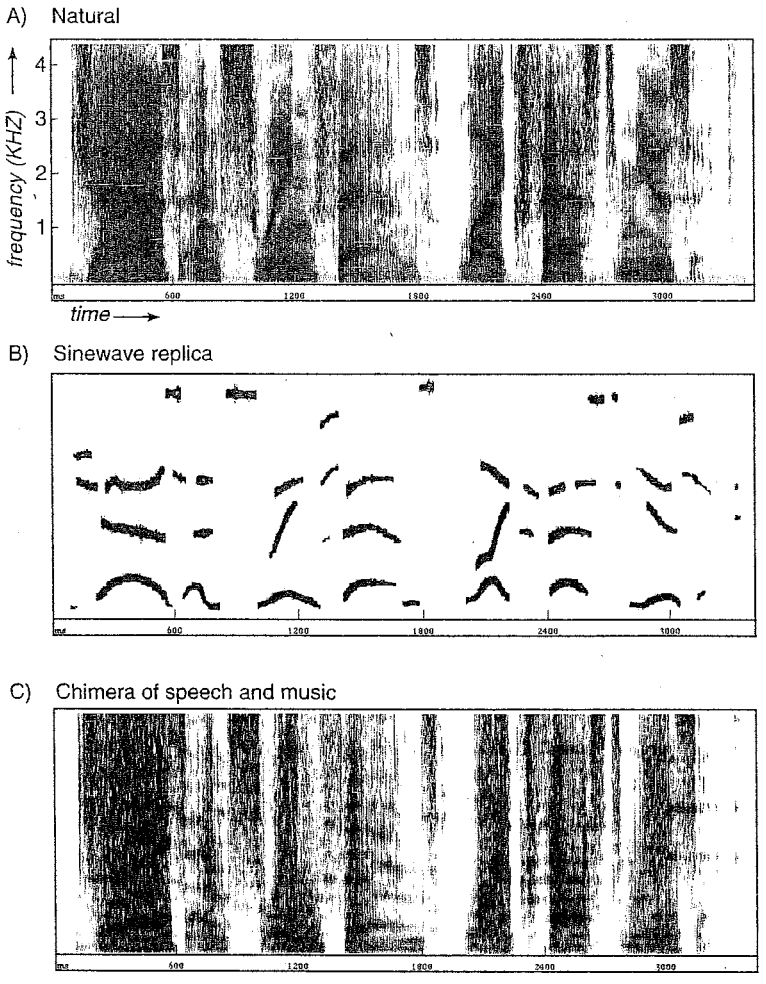

Figure 1. Inventory of acoustic constituents of three kinds of intelligible sentence: (A) natural speech; (B) synthetic replica composed of time-varying sinusoids; and (C) acoustic chimera made by exciting the changing spectrum envelope of speech with the fine structure of a nonspeech sample recorded by Count Basie in 1939.

1.1.5. Nonsymbolic

A study using patterned sinewave lures revealed that perceptual organization is nonsymbolic in its effects. That is to say, perceptual organization occurs by sensitivity to speechlike variation, and is distinct from the analytical finesse that creates an impression of linguistic form (Remez, 2001). In this test, the components of a sinewave sentence were arrayed dichotically, separating the tone analog of the first, third and fourth formants in one ear from the analog of the second formant in the other. The challenge to organization was to resolve the coherent variation among the tones composing the sentence despite the spatial dislocation of its constituents. After establishing that listeners tolerate spatial dissimilarity in amalgamating the tones perceptually, a variety of lures was introduced in the frequency band of the second formant in the ear opposite the true tone analog of the sentence, which also contained the tone analogs of the first, third and fourth formants. In this kind of presentation, perceptual organization is challenged to resist the lure presented in the same ear as the first, third and fourth tone, and to appropriate the second tone that completes the sentence. Some lures were easy to resist; those that were constant in frequency or those that alternated brief constant frequency tone segments did not harm the organization of dichotically arrayed components. Other lures were far more difficult to withstand: a gradient of speechlike variation was created by straining (or, more commonly, squashing) the frequency variation of the lure to vary from speechlike to constant frequency at the average frequency of the second formant; the speechlike variant was actually a temporally reversed second formant analog. The lure could not complete the sentence because its variation was never coherent with the first, third and fourth tone. Nor did it evoke impressions of linguistic form; the lure in the clear sounded like a warbling pitch pattern. Nonetheless, it interfered with organization in proportion to its frequency variation, the most when it varied with the pattern of a second formant; less when it varied in frequency at the pace and pattern of a second formant but over a reduced frequency range; and not at all when it was squashed to a constant frequency. In other words, the propensity of the lure to interfere with perceptual organization depended on the speechlike properties of its frequency variation alone, because this single time-varying tone did not evoke phonetic impressions that competed with those of the sentence.

1.1.6. Requires attention

A subjective impression of inevitability often accompanies the perception of utterances, but this belies the actuality; the perceptual organization of speech requires attention, whether elicited or exerted. This is seen, again, in studies of sinewave replicas in which the auditory quality is so little like speech listeners are unlikely to organize the tone complexes spontaneously (Remez et al., 1981; Remez, Pardo, Piorkowski, & Rubin, 2001). When a sinewave replica of an utterance is heard simply as a set of contrapuntal tones, none of the auditory qualities is vocal, hence nothing about the experience compels the perception of linguistic properties. Physiological measures were consistent with the hypothesis that no covert aggregation of the tones occurs if the perception of auditory form takes place without apprehension of phonetic attributes (Liebenthal, Binder, Piorkowski, & Remez, 2003). However, a listener who is informed that the tones

compose a kind of synthetic speech is readily capable of transcribing a sinewave sentence on this instruction alone; no training, extensive exposure or special hints are needed. In this condition, the aim of hearing the tones phonetically permitted attention to the coherence, albeit abstract, of the speechlike variation in the tones. Moreover, only the sinewave patterns derived from natural utterances are amenable to organization by virtue of the exercise of attention. An arbitrary or incomplete physical spectrum does not evoke an impression of phonetic attributes simply because a perceiver intends to resolve a speech stream. Because natural and much high-quality synthetic speech elicits phonetic attention by virtue of intrinsically vocal auditory qualities, the role of attention in speech is easily overlooked. More generally, this finding of a contingency of organization on attention in the case of speech anticipated the claim that auditory perceptual organization in the main depends on attention (Carlyon, Cusack, Foxton, & Robertson, 2001). It also falsifies the claim that speech perception is accomplished by a modular faculty, because the contingency of perceptual organization on attention contradicts the premise of autonomous and mandatory action immune to influence by belief.

1.2. Audiovisual Perceptual Organization

The classic formulation of perceptual organization took the cocktail party as the critical setting, and much of the ensuing research examined the proficiency with which a perceiver pulls the speech of a talker of interest from a lively acoustic background. Although this conceptualization has been productive, a different slant is needed to describe a listener who can also see the talker. It has been well established that in this situation a perceiver treats speech as a multimodal event, sampling visually and auditorily (Sumby & Pollack, 1954). In multimodal speech perception, the formal characterization of finding and following a speech stream remains much the same as the auditory instance, with a twist. Rival conceptualizations have characterized multimodal perceptual organization in parallel streams converging at the end or, as a single multimodal stream in which visible and audible features interact continually. To caricature the two perspectives, the first pictures the perceiver as both a blind listener and a deaf viewer huddling within a single skin, resigned to negotiate any discrepancy between utterances that each determines after perception concludes. The second perspective conceives of the perceiver as an auditory-visual synesthete in whom visible and audible ingredients blend so thoroughly from the start that no residue remains to distinguish the sensory core of the phonetic forms. Whichever conceptualization comes closer to the truth, the visual resolution of speech poses the familiar challenge to organization: the physical effects are regular albeit unpredictable, and none of the effects in detail is unique to speech. This is so whether the level of description is a stream of light reflected from the surfaces of the face, a 2 1/2-d sketch of an as yet unresolved face in a visual scene, or a description of the face as a familiar kind of object in motion.

1.2.1. Intersensory combination

In considering the problem of multimodal perception, it is also natural to speculate about the grain of analysis at which intersensory combination occurs (Rosenblum,

2005; Lachs & Pisoni, 2004). Principally, this is a puzzle to solve only if the second alternative conceptualization, of the amalgamated multimodal stream, proves to be true. If speech perception occurs separately in visual and auditory modalities, then perceptual organization proceeds in parallel for visible and audible samples, and any interaction between modalities occurs after perceptual analysis has resolved the phonemic form. Because the phonemic properties are set by their contrastive linguistic function, and not by the specific transient sensory or motor forms of their expression, the dimensions of visually perceived speech intrinsically match those of auditorily perceived speech. Under this condition, the scientific puzzle of understanding the alignment of visible and audible streams is obviated.

The opposite conceptualization, in which intersensory blending occurs in organization, poses a puzzle, for there is no obvious dimension common to vision and hearing. The sensory qualities of these two modalities are largely incommensurate – hue, brightness and saturation do not form tight analogies to pitch, loudness and timbre – and it should be evident that the acoustic transparency permitting a listener to hear the changes originating in the action of the glottis and tongue body have no counterpart in a visible face, in which the larynx and all but the lips and the tip of the tongue are out of sight. Some research proposes the existence of a common intersensory metric, an intermediary permitting the visual and auditory streams to blend in a form exclusive to neither sense (Massaro & Stork, 1998). Variants of this proposal cast auditory sensation – pitch, loudness and timbre – as the common metric into which visual form is also cast for a spoken event (Kuhl, 1991), and a kind of shallow representation of visual and auditory primitives in articulatory parameters (Rosenblum & Gordon, 2001), about which there is much more to say when we turn to phonetic analysis.

Evidence favoring each conceptualization of intersensory relation exists in the technical literature, chiefly in studies of audiovisual merger. That is, synthesis and digital editing of video and audio components have been used to create phonemically discrepant visible and audible presentations, with which to determine the nature of multisensory combination. In the original use of this method (McGurk & McDonald, 1976), an audio [bɑ] and a video [gɑ] were resolved as a fusion, [dɑ]. In tests of this kind, it is possible to fix the identifiability of auditory and visual components independent of tests of their combined effect, and such findings are subsumed well within a parallel model of perceptual organization in which concurrent perceptual states are reconciled if the perceiver organizes them as bound to the same talker. Alternatively, some audiovisual phenomena are simply not well described by parallel and segregated organization in each modality.

In one of these intriguing cases, a video of a face was presented with an amplified electroglottograph signal correlated with the pulsing of the larynx of the depicted talker (Rosen, Fourcin, & Moore, 1981). The appearance of the face was unexceptional; the electroglottograph sounded like an intermittent buzz changing in pitch in the range of the voice. Some of the syllables and words could be resolved phonetically by watching the face, and the audible buzz evoked no impressions of words at all. Overall, the

conditions for unimodal visual speech perception were barely met, and were not met at all for unimodal auditory speech perception. In this circumstance, multimodal perception should be poor in as much as the cumulative effect of poor visual and no auditory perception remains poor. Instead, the combination was fine, arguably reflecting the effectiveness of auditory and visual streams in combination when separately neither stream was adequate. This finding among many others offers evidence of a common dimensionality for viewed and heard speech preliminary to analysis.

1.2.2. Mismatch tolerance

Among the best clues to the nature of multimodal perceptual organization are the results of studies of the tolerance of spatial and temporal discrepancy across the modalities. In one notion, vision and hearing supply discrepant but complementary samples of speech (Bernstein, Auer, & Moore, 2004), and this simplification describes both audiovisual presentations in which visible and audible patterns coincide and cases of intersensory competition. The organization of fine grain discrepancy between viewed and heard speech scales up to coarse grain discrepancy, and this functional similarity over scale variation is surprising. At the finest grain, auditory and visual streams are mismatched simply because of the disparity in the aspects of the physical acts of articulation that each provides, and not only because the primitives of auditory and visual sensation differ. Indeed, the enterprise of multimodal research rests on findings that fine grain discrepancies introduced by a scientist's method are resolved in perception, often without eliciting an impression of disparity in the seen and the heard speech. And, at the coarser grain as well as the finer, perceptual integration of discrepant sensory samples is robust.

In several studies, perceptual integration survived spatial displacement of audio and video sufficient to notice (Bertelson, Vroomen, & de Gelder, 1997); and, temporal misalignment of audio and video sufficient to notice (Bertelson et al., 1997; Munhall, Gribble, Sacco, & Ward, 1996). To be more precise, the merging of phonetic features used to index the perceptual integration of vision and hearing persisted under conditions of sizeable spatial and temporal divergence. Such findings can leave the researcher without a convenient explanation because the theory of first resort fails to apply. Specifically, the very tolerance of mismatch blocks the psychologist's automatic and tiresome appeal to similarity as the engine of integration; the integrated streams are dissimilar, displaced and lagged. And, the conditions created within the audiovisual display introduce discrepancies at a scale that surpasses ordinary experience by an order of magnitude. Appeals to likelihood can seem clichéd in psychological explanation, but this procrustean tactic must fail in these instances. The relative divergence of the integrated streams is just unfamiliar.

1.2.3. A unimodal and multimodal contour

How, then, does a perceiver apprehend the disparate sensory samples of speech as a coherent progressive event? When organization is veridical, the auditory or visual effects are grouped despite dissimilarity and discontinuity of the sensory constituents of a

perceptual stream. The familiar principles of perceptual organization deriving from Gestalt laws of figural organization (Bregman, 1990) cannot be responsible for unimodal organization, for they invoke one or another form of similarity among the sensory constituents. If a role for this conventional account seems unlikely to suffice in unimodal organization, it is utterly implausible for explaining the cases of multimodal organization in which some form of binding appears to occur intermodally in advance of analysis. Although the discussions in the technical literature generally portray binding as a process of sorting analyzed features into bundles coextensive with objects, it appears as though the urgency of auditory perceptual organization compelled by the fast-fading sensory trace imposes a different order. Instead, binding of the sensory constituents of the spoken source must occur before analysis, and perhaps this is the cause of the condition that the perceptual organization of speech requires attention. It is tempting to speculate that there is a single set of organizational functions that applies regardless of the assortment of samples arrayed across the modalities, and some studies of neural metabolism correlated with perceptual organization (Liebenthal et al., 2003) are consistent with this view – but evidence of its existence holds far less value than evidence of its characteristics would.

1.3. Perceptual Analysis

A perceiver who resolves a stream of speech in a raucous or tranquil scene might also be able to resolve its linguistic form. These facets of perception are contingent. Certainly, the circumstance in which a listener knows that someone is speaking but cannot make out the words is familiar to us, although the inverse – linguistic impressions of a spoken event in the absence of an impression of someone speaking – might merit a thorough reappraisal of mental status. The perceptual resolution of linguistic form has been a topic within the technical study of speech for more than seventy years, and the longevity of this concern is due to the intriguing complexity of this type of sensitivity. Although it has taken a variety of guises, in each the central challenge has been to understand the perceptual ability to apprehend the expression of a small number of linguistic forms under conditions that vary without end.

Long ago, research on the perceptual analysis of speech adopted a focus on the ultimate constituents of language. That is, the linguistic properties that speech expresses are componential, and the components are hierarchically nested. Utterances in the form of sentences are composed of clauses, within which phrases are nested; phrases comprise words, words are composed of syllables and each syllable can be a series of phoneme segments. Phonemes are grouped by distinctive features, that is, by virtue of the coincidence of disjunctive attributes that, together, constitute a system of contrasts across the segmental inventory of a language.

To researchers of the first generation of psycholinguists, the componential nature of linguistic structure was theoretically significant, though the focus on ultimate constituents in speech perception research was also practical (Miller, 1965). It is not sensible to focus on sentences as irreducible objects of perception— there is an infinite

number of them, and of phrases, too. Although languages differ in the number of words that they sustain at any moment in history, the set of these is also large. To consider a specific instance, in English, a language often studied in psycholinguistics, words derive from Germanic and from Romance heritages, and for this reason English is said to incorporate more words than is typical of other languages with calmer history. A focus on individuals of a specific group can restrict the study to vocabulary in regular spoken use – in contrast to the far larger recognizable vocabulary – which imposes an inventory of roughly 15,000 items (Miller, 1951). If this is a saving from the infinity of sentences and a large lexical stock, even greater economy is achieved by considering that whatever the word, in English it is composed from a supply of three dozen or so phonemes expressing perhaps a dozen and a half contrast features. Taxonomies of phonemes and the features on which the classes are sorted can become controversial from time to time, depending on the rise or fall in value of one or another kind of evidence. Even with such disputes, there has been good agreement that the perception of speech entails the perceptual resolution of elementary linguistic attributes available in a brief spoken sample; larger structures of linguistic form are produced cognitively by aggregating the elementary constituents provided by speech perception. We defer a discussion right now of the consequences of the phonetic expression of phonemic contrasts, but not for long.

Setting a perceptual focus that is linguistic, segmental and contrastive defines the products of perception, although consensus about the effects has not tempered the disagreements about the causes of perception. This dispute among perspectives concerns the kind of perceptual analysis yielding the linguistic objects. Proponents have divided on its essential nature. Either the perception of speech depends on auditory sensitivity and categorization, or on articulation, or on linguistic function. Each of the proposals is old, and the stalemate is apparently perpetual. We will offer a recommendation, but first we expose some of the technical details.

1.3.1. A general auditory account

The roots of the auditory approach run deep. Among the earliest reports in experimental psychology are studies of the likeness of simple whistles and buzzes to speech sounds (Kohler, 1910; Modell & Rich, 1915). Although the correlations were only rough, they licensed the claim that *vocality* is a primitive auditory sensory quality. The argument held that because vocal impressions are elicited by simple acoustic attributes they are fundamental in human sensory experience; therefore, a talker's ability to evoke a listener's phonemic states depends on producing sounds that hit auditory targets given subjectively and intrinsically. There are more technically sophisticated versions of this antique claim at large today (Kuhl, 1991), but the germ of the idea is similar. Indeed, such findings are perennially welcome in psychology due to a resilient eagerness for sensory reduction of perceptual impressions of the structures and motions of the world. The draw of this explanation is that it permits a description of perception to attribute an incidental role to the objects and events that ultimately cause sensory states: All of the explanatory action pertains to the sensory pathway and neural centers of associative learning. Indeed, it has become commonplace recently for the justification to invoke a perceiver's ability

to learn the statistical characteristics of the distribution of sensory states with which phoneme contrasts allegedly coincide. This premise invokes a hypothetical norm in its attempt to accommodate the variability in the acoustic form of each phoneme due to the variety of talkers, rates of speech, and attitudes expressed concurrent to language production, each of which precludes an acoustically uniform expression of a phoneme across different occasions. In one expression of this idea (Diehl, Kluender, Walsh, & Parker, 1991), the auditory system is viewed as a nonlinear conduit of the acoustic effects of speech in which contrast is created by means of enhancement of some auditory elements relative to others. Admittedly, adherence to a general auditory perspective is only weakly justified by psychoacoustics or auditory physiology (Diehl, Lotto, & Holt, 2004).

The perspective on speech perception offered in a general auditory approach has a goal, to pursue a model of the phonemically interested listener as a trainable ear and little else. In a recent review, Diehl et al. (2004) argued that the explanatory detail presently accrued under this rubric is too thin to permit a falsifying test, but this reservation seems unduly gloomy. Even if precise predictions of experimental findings are not readily produced from the principles underlying the approach, it is sensible to ask if the premises of the model attach importance to false assertions. Specifically, if the ambition of the model is not mistaken, its allure is surely diminished by two well established properties of speech perception: (1) the fleeting nature of auditory forms; and, (2) the irrelevance of auditory norms. First, in this class of accounts, perception is based on the varying sensory correlates of speech sounds, and a listener's personal history of experience with /d/, for instance, is encoded to generate a long-term probability distribution in which more and less typical auditory manifestations of /d/ are calibrated. The success of a listener in recognizing instances of this phoneme would necessarily depend on the likelihood that an as yet unidentified sensory form can be assimilated to a longstanding likely auditory representation of /d/ among other segments in the language. But, classic psychoacoustic research revealed that the auditory properties of speech are exceedingly fragile, and are difficult to protect for even a quarter of a second (for instance, Howell & Darwin, 1977). This limit must be a mild embarrassment, at least, to a conceptualization relying on the durability of raw auditory impressions of speech. Although such representations are reasonably chosen for instrumental applications such as speech-to-text systems, these are constrained by circuit design and not by physiology (Klatt, 1989). To survive in a listener's memory, short-lived auditory properties acquire a different form, possibly in the dynamic dimensions of the sources that produced them (Hirsh, 1988), and when a listener remembers a sound, it is more likely that the recalled quality is generated rather than replayed from a faithful inscription in memory of the original auditory form.

A second problem for a general auditory account of speech perception is its reliance on auditory manifestations of the phoneme contrasts graded by likelihood. Even to entertain this premise, we must be credulous momentarily about the prior claim that unelaborated auditory forms of speech are retained well in memory; this suspension of criticism permits us to review the assertion that a spoken phoneme is identified by a normative assessment of its sensory form. In short, the robustness of intelligibility over widely varying natural conditions of acoustic masking and distortion show clearly that

neither goodness nor typicality in auditory quality is requisite for speech perception. Indeed, intelligible sentences are perceived from patterns dissimilar to speech in acoustic detail and in auditory effect (Remez et al., 1981; Shannon et al., 1995; Smith et al., 2002). But, what is the shape of a distribution of the auditory attributes of a phoneme?

To be truthful, no one knows. There is a single study of actual incidence, of the exposure of a single infant to speech produced by one adult (van de Weijer, 1997). This means that claims about sensitivity reciprocating the distributions of the acoustic or auditory forms of speech are hopeful, and without empirical foundation (Saffran, Aslin, & Newport, 1998); at least, no claim is grounded empirically yet. But, it is not difficult to recognize the implausibility of the claim that auditory typicality determines the perception of phonemes. The typical auditory forms of speech must be those sensory states evoked by exposure to the acoustic products of vocalization. After all, an overwhelming majority of instances must be those in which a listener perceives speech because a talker spoke. These days, the pervasiveness of the experience of speech over the telephone also contributes to normative distributions, and so does speech produced by talking toys and gadgets. Overall, the probability distribution must represent this kind of typical experience composed chiefly of acoustic vocal products with minimally distorted variants at the improbable ends of the distribution.

In fact, listeners are evidently not fussy about the acoustic constituents or the auditory qualities of intelligible signals. Neither natural broadband resonances nor harmonic excitation nor aperiodic bursts and frictions nor any specific set of acoustic correlates of a phoneme is required for perception (see Figure 1). Instead, a listener perceives speech as if the commitment to the particular sensory realization of the linguistic contrasts is flexible. This readiness to find functional contrasts in the least expected acoustic or auditory form opposes the fixity of an audit ory norming rationale. Indeed, such acoustic norms – some without auditory warping – form the basis of speech-to-text devices often aimed at the typical expressions of just a single individual (Picheny, 2003); even so, we are still typing.

Before turning to consider an account of perception grounded in articulation, it is useful to note that there are important questions about auditory function in speech that do not depend on the claim that phoneme categories coalesce out of auditory form. At the most elementary, the acoustic correlates of each linguistic contrast are multiple: the speech stream itself is a composite of dissimilar acoustic elements. Attention to the auditory quality of constituents of a speech stream – an aperiodic burst, a second formant frequency transition, a noisy hiss – can occur concurrently with attention to the linguistic properties – an unvoiced fricative of coronal articulatory place. This kind of *bistable* perception in which attention can hold auditory form or its superordinate, or both, is not well understood outside of musical contexts. Moreover, if qualitative attributes of speech are retained in a durable form, the dimensionality of such knowledge is not well explored (Hirsh, 1988). At the largest grain, the flexibility of the standards for perceiving the linguistic elements of speech is well evident, yet the function by which a perceiver resolves linguistic properties in specific instances, especially those evoking novel auditory form, remains a tough puzzle.

1.3.2. An articulation-based account

Modern linguistic description took shape with phoneme contrasts already described in the dimensions of articulation. The technology required to portray acoustic properties did not exist, and in the resort to articulatory dimensions to describe the sounds of speech, Joos (1948) says linguists made a virtue of necessity. However, this practice was unsatisfactory even as articulatory description, largely because the method presumed anatomical and functional states of articulators without direct evidence. For instance, the classical notions of articulatory contrasts in vowel *height* and *advancement* were designated by intuition, not by observation, and ultimately proved to be inaccurate portraits of the tongue shape and motion discovered in x-ray fluoroscopy, electromyography and magnetic resonance imaging (Honda, 1996).

When methods for direct measurement of sound became available to supplement impressionistic descriptions, it had a paradoxical effect on the restlessness with old fashioned articulatory description. As the basic properties of speech acoustics were described technically, a problem emerged for proponents of acoustic description; indeed, the conceptualization of articulation was challenged as well. Research on production and perception alike failed to find counterparts to the theoretical description of phonemes in articulatory, acoustic or auditory components. Each perspective in its own way had presupposed that speech was a semaphore, with every phoneme a kind of vocal act or pose, or every segment a kind of acoustic display. Instead, whether construed as acts of articulation, their physical acoustic products or their psychoacoustic effects, an apparent lack of invariance was evident in the correspondence of the linguistically contrastive phoneme segments and their expressive manifestations. In each domain, the relation of a phoneme to its articulatory and acoustic correlates proved to be one-to-many.

Of course, the mere existence of variety among the physical or physiological correlates of a linguistic component is not troublesome to perceptual explanation. If the articulatory, acoustic or auditory tokens of different phonemes correspond uniquely to types, the lack of invariant form is insignificant because the correlates of one type are not shared with any other. The critical finding about the relation of phoneme to correlate was the nonexclusive relation between type and token. One of the clearest instances is the /pi/-/kɑ/-/pu/ phenomenon (Liberman, Delattre, & Cooper, 1952) in which a single acoustic element evokes an impression of a *labial* consonant and a *palatal* consonant depending solely on the vowel with which it is presented.

A key explanatory innovation occurred in response to such findings. A new sense of the idea of *coarticulation* was created to describe the relation of production and perception; a history of coarticulation in phonetic linguistics is offered by Kühnert and Nolan (1999). At the heart of this breakthrough was the inspiration that descriptively segmental phonemes are encoded in articulation (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). That is to say, a perceived phoneme sequence is restructured rather than produced as a simple sequential cipher or articulation alphabet. The encoding occurs because the vocal articulators are intrinsically separable into controllable parts, and the expression of a sequence of phonemes is thereby reassembled as an imbricated pattern of

constituent acts that unfold concurrently and asynchronously. This approach explained well the inexhaustible variety of articulatory and acoustic correlates of each phoneme, or, rather, the lack of a consistent physical manifestation of a phoneme, because whichever segments preceded and followed it shaped its articulation by contributing to the encoding; and, no phoneme is ever expressed in isolation of coarticulatory influence. Liberman et al. explain that such recoding achieves high rates of segmental transmission with sluggish anatomy. The cost is to obscure the relation between an intended or perceived phoneme and its articulatory and acoustic form. Accordingly, speech gives phonetically encoded expression to an intended if abstract phoneme series.

From this premise, a characterization of perception follows as directly as night follows day. If the acoustic speech stream is an encoded version of phonemes, due to the articulatory restructuring of an abstract segmental series, then an inverse operation required to apprehend the segmental series obliges a perceiver to reciprocate the motor encoding in some fashion. A variety of specific technical hypotheses about this kind of perception was ventured in different versions of the *motor theory*, including reliance on learned articulatory correlates of the auditory forms of speech; covert efferent mimesis; and, imagined surrogates of proprioception that accompanied a talker's speech.

Challenged by evidence, the motor theory looked terrific at a distance, from the perspective of neuropsychology or studies of human evolution (Galantucci, Fowler, & Turvey, in press). At close range, the disconfirming proofs of its technical claims emerged steadily from detailed research on the relation between perception and production. Crucially, studies of extremely young infants showed that perceptual sensitivity develops in advance of articulation, and is not a consequence of it (Jusczyk, 1997). In adults, the invariant characteristics presupposed of the articulation of individual phonemes was falsified in studies of articulatory motion and electromyography (MacNeilage, 1970). This is an enormously intriguing literature impossible to gloss. Yet, acknowledging exceptions, the fair preponderance of evidence showed that every phoneme takes many anatomical forms, and invariance in the correspondence of phoneme to motor expression was found neither in an aggregate of α-efferent activity, nor in the precise motion or configuration of articulators, nor in the shapes of the vocal tract achieved by articulation. In a revision of the motor theory proposed to answer research that it had motivated, perception was held to resolve the invariant phonemic intentions of a talker rather than the acts of articulation, varying without limit, as they are executed (Liberman, & Mattingly, 1985). This version represents spoken communication as a transaction composed of deeply encoded phonemic intentions, aligning the revised motor theory with the symbolic emphasis of a linguistic view of speech perception.

A pair of conjoined accounts of more recent vintage aims to span the gulf between intention and action while retaining the emphasis on production of the motor theory: *articulatory phonology* and *direct realism* (Goldstein & Fowler, 2003). Articulatory phonology offers a description of linguistic contrast set in abstract articulatory primitives, and direct realism describes a perceiver's sensitivity to articulatory contrasts by attention to the visible, audible and palpable effects of speech. The crucial contribution of this

proposal is a representation of lexical contrasts in a repertoire of *gestures*, not of phonemic segments. Building on the characterization of articulation given by Liberman et al. (1967), a contrastive gesture is designated as: (1) a movement of a particular set of articulators; (2) toward a location in the vocal tract where a constriction occurs; (3) with a specific degree of constriction; and, (4) occurring in a characteristic dynamic manner. In this perspective, a word is indexed by a gestural score describing its production as the coupled asynchronous action of lips, tongue tip, tongue body, velum and larynx. Such gestural components are understood as quasi-independent actions of vocal articulators. The pattern with which gestures impose and release constrictions creates the contrasts customarily described in a segmental phoneme series. The traditional separation of phonemic contrast and phonetic expression theoretically collapses in this account into an equivalence between linguistic properties and vocal acts. With respect to the principle at the core of the motor theory, this asserted equivalence of linguistic contrast and manifest articulation denies the encoding that supplied the articulatory character of the inverse function purportedly applied by a perceiver to a speech stream. In complementary function, the account describes the perceiver distinguishing words in the same gestural components that the talker employs to create speech. The phonemic properties of speech are apprehended perceptually without decoding them, according to this argument, because the acoustic pattern and its sensory effects are transparent to the articulatory components that index spoken words across the lexicon (see Figure 2).

Part of the attraction of this theoretical gambit is the potential of an articulatory phonology to explain allophonic variation in a natural way. That is, the production of a

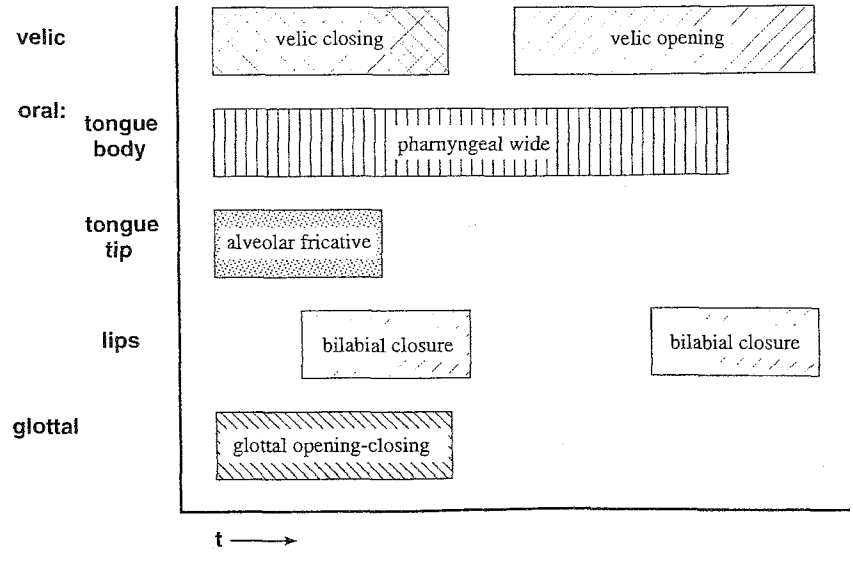

Figure 2. A gestural score representing the coupled actions of quasi-independent vocal articulators in the production of the word SPAM. (Browman & Goldstein, 1991, p. 318).

canonical phoneme sequence actually varies in exact phonetic, or expressed, detail. The word SECURITY, for example, is produced as these variants, among others: [sək^{h}jʊɹɪt^{h}i], [sɨk^{h}jʊɹəɾi] and [sk^{h}jəˑɾi]. Under an articulatory phonology, many variants are potentially rationalized as consequences of minimally different task dynamics of the same gestural components given in a lexical representation. Variation attributable to differences in speech-rate, reductions, lenitions and apparent deletions are likewise described as natural variants of the same gestural form, and this sameness remains available to perception, in principle. A listener who resolves the gestural components in a speech stream might notice but pay little attention to the effects of slight phase differences in the expression of constituents of an intended contrast. Even in casual speech, in which the canonical forms of words can be compromised by the expressive aims of a talker, articulatory phonology promises to explain the variation without representing the phonetic form of the articulation differently from the phonemic form distinguishing a word from all others. Admittedly, in some synchronic and diachronic cases, it seems that talkers do express different contrasts than the lexicon employs, yet the different representations warranted by these facts can nonetheless be described by postulating no more than minimal changes in the components of a gestural score. Of course, there are some phonological phenomena in some languages that defy simple characterization – Nature provides her own exceptions – and it is not clear how these will be resolvable to general principle in relating the phonemic and expressed forms (Browman & Goldstein, 1991). But, the perceptual claims of this account are readily evaluated.

Two critical axioms are assumed in the perceptual account given by articulatory phonology and direct realism, and if they are not exactly false, they are less true than the account demands. The first is an asserted *isomorphism* between the components used in language to create contrast and the components of spoken acts; they are designated as a single set of gestures. The second is a state of *parity* such that talker and listener match; the expressed forms and the perceived forms are claimed to be the same. Of course, these axioms are related. People who speak the same language use the same canonical contrasts. If they express them differently, or if a talker nonaccidentally expresses the same form in gestural variants on different occasions, then the relation between canonical and expressed forms can be regulated, adjusted or reshaped; phonemic and phonetic form are not identical in this circumstance. Instead, some of the degrees of freedom in articulation would be reserved for expression beyond those that are committed to the canonical form of the word. If articulation varies with a talker's communicative aims, then canonical and expressed forms do not match, and parity must be achieved, not simply fulfilled.

Is articulation isomorphic to phonemic contrast? A recent set of examples described by Hawkins (2003) shows the extent of the mismatch in ordinary circumstances. The graded expressions of the utterance, "I do not know," that she discusses range widely: [ɑi du nɑt nɷu], [ɑi dɷnt nɷu], [dənɷ], and [ə̃ə̃ə̃]. If this shows the varieties of phonetic form under different expressive demands, consider an example of the same expressive demand imposed on different individuals. Three adult talkers from Southern New England contributed to a corpus of sentence-list reading (Remez, Fellowes, & Rubin, 1997), each producing the word DROWNING in consistent yet distinct form [dɹɑʊnɪŋ], [dʒɹɑʊnɪŋ] and

[dsɹɑʊnɪŋ]. Even a single individual in a stable pragmatic condition can generate such variation. In a recent observation of a plausibly commonplace incident, we witnessed a taunting child calling her father: [dæɾi]... [diæɾi]... [dædi]... [dæːɾi]... [dæʔ di]... [dæɾ̩i]..., etc. In each of these examples, it is difficult to maintain that phonetic form is well described as an obligatory isomorphic projection of canonically given gestural form. While these cases obviously express phonemically contrastive gestures under mutual coarticulatory influence, they also show the converging influence on phonetic form of long-term effects of dialect and idiolect, and the immediate effect of pragmatic pressure and opportunity. In the causes of departure from isomorphism, there is more to explain than the effects of coarticulation and rate variation.

Can a phonology indicate the aspects of the phonetic realization that stem from canonical lexical form and those that are controlled by consistent non-lexical aspects of expression? Surely, dialect and idiolect are components of language, and the phonetic means of marking contrasts between communities and among talkers within them shares the phonetic grain of production with the expression of lexical contrasts. Perhaps an ultimate model would include the influence of long-standing or momentary expressive aims on the articulatory parameters in the task dynamic organization of production, explaining the regular expressive variants of canonical lexical form that supersede straightforward coarticulatory and rate effects. But, the consistent lapses in isomorphism between canonical and expressed form deprives the perceptual account of its central claim: the transparency of lexical contrast sensed via public aspects of speech. In the absence of isomorphism, the perceiver is challenged to resolve the canonical gestures from a speech stream that marks dialect, idiolect, affect and attitude in phonetic form as well. Variation in phonetic form that disrupts simple isomorphism with phonemic form is often conceptualized as noise, the consequence of undeniable but unpredictable deflection from canonical expression, but this is pessimistic and mistaken (for instance, Pisoni, 1997). Perceivers are sensitive to dialect variation even if intuitive dialect geography is no more accurate than intuitive geography. Attributes of idiolect independent of vocal quality are evidently tracked and integrated perceptually, and are effective for identifying individuals by discerning their phonetic habits (Remez et al., 1997). The phonetic basis for perceiving affect and attitude are harder to pin down, empirically, but in tractable cases like smiling – a gesture of the lips – the phonetic effects are sensed readily (Tartter, 1980). In other words, a perceiver is often capable of resolving the aspects of phonetic expression that derive from its multiple causes, and a complete account of perception would aim to describe the facility with which we attend to the properties of articulated linguistic form that count lexically and the collateral resolution of those that are consistent and informative but not distinctive lexically.

The axiom of parity denotes sameness in language forms shared by a talker in composing an expression and by a listener in perceiving it. Here, the intended sense of parity applies only to the gestural components of language, and this is completely apt. After all, lexical parity is commonly breached: conversations are fine with people who say LIGHTNING BUG when we say FIREFLY, SACK for BAG, TRASH for GARBAGE, that is, under conditions in which the communicative function matches while the lexical form does not.

The claim of parity states that the forms of perception and production are the same, and the claim at some level of resolution cannot be false (Liberman, 1996). If the perceiver knows what the talker said well enough to repeat it, lexical and phonemic parity occurred, at least. But, because isomorphism is suspicious, expressed forms can be understood to differ from abstract phonemic forms. The axiom becomes harder to sustain in that case because phonemic parity can occur without phonetic parity; and, because phonetic parity is so unlikely, even in monozygotic twins reared in the same household (Johnson & Azara, 2000; Nolan & Oh, 1996; cf. Gedda, Bianchi, & Bianchi-Neroni, 1955).

Critical data on this topic indicate that perception is bistable, permitting attention to be drawn to superficial and canonical form concurrently. The study (Goldinger, 1998) used an original measure of perceptual resolution. An experimenter presented a recorded utterance for a subject to repeat immediately or after a brief imposed delay. Comparing the elicited speech samples to the eliciting sample showed that utterances produced immediately were more similar to the eliciting sample than were those produced after even a brief delay. Despite all, a similar utterance was far from a faithful replica of the model. This is expected, to be precise, for not even nightclub impressionists achieve their characterizations by exact replication of the speech of John Wayne and Cary Grant – and they rehearse. The difference in the two conditions of lag must be attributed to phonetic attributes inasmuch as the words were the same, hence, the contrastive phonemic properties were the same. With respect to the parity axiom, though, the result is troubling, because the finding of only rough similarity insinuates that if parity is fostered it is unattained; and, that the faint shadow of parity that actually is manifest lasts only a moment, and once the impulse toward parity subsides the default state of disparity returns. Moreover, studies of deliberate imitation show that an individual typically provides an erroneous imitation of a self-produced speech sample (Vallabha & Tuller, 2004). If parity does not occur in this limiting case, when would it?

A plausible description of these phenomena is possible without invoking a disposition to isomorphism and parity. In the moment when a spoken word is perceived, phonemic and phonetic forms are resolvably different from each other. The salient differences often include aspects of a talker's speech that differ greatly or minimally from a perceiver's own characteristic articulation. Speech initiated in this state can be nudged toward the form of the immediate phonetic model and away from the habitual expression of canonical form. At a greater delay, though, the vividness of the phonetic impression of the eliciting utterance has faded in its contrast with long established articulatory habits, and production is free from the adulterating pressure of a phonetic form distinct from the talker's intrinsic dynamic. It is as if talker and listener express lexical and phonemic parity by means of their phonetic differences. With sustained exposure to an individual talker, a perceiver is likely to form an impression of the talker's characteristic articulatory variation, sufficient to imagine speech produced in the voice and style of the talker, and perhaps to adopt the phonetic characteristics in a deliberate imitation (Johnson, Foley, & Leach, 1988). There is some evidence that such vicarious experience of the speech of familiar others can influence a talker's production in detail (Sancier & Fowler, 1997). But, the listener and talker need not match phonetically for any of this to occur. Indeed, in order for the phonetic similarity of two talkers to wax and wane, they cannot match.

The assertions of isomorphism and parity mask a significant aspect of the perception and the production of speech, namely, the ubiquity of mismatching form. Whether the discrepancy occurs in the visible and audible properties of speech, as in audiovisual speech perception, or in the phonetic realization of phoneme contrasts, as occurs whenever two individuals speak to each other, it seems that you neither expect nor require your conversational partners to use the identical expressive forms that you use. Or, more precisely, the sharing of words apparently licenses variegation in articulation, both in groups – as dialect unless the group also possesses an army and a navy, in which case it is a language – and in individuals – as idiolect.

1.3.3. Perceiving speech linguistically

A linguistic emphasis in explanations of speech perception is familiar. The basic notion deriving from Jakobson and Halle (1956) identifies phoneme contrasts as symbolic and linguistic, and neither articulatory nor auditory. In this regard, they assert symbolic status to the phoneme and the word alike. This is subtle, for it warrants a distinction between the form of words ("I said PIN, not PEN") and their meanings ("I meant PIN, not PEN"). The relation between sound and meaning is arbitrary notwithstanding the contrary claim of *phonesthesia*, a perennial topic of romantic symbolists (Aman, 1980). In order for the listener to know what the talker meant, the listener must resolve the form of the talker's utterance; without grasping the form of a talker's speech, a listener has merely guessed the talker's meaning. It is this juncture that is critical for this conceptualization, because of the complexity in the relation between the canonical form regulated by the language and the expressed form regulated in compromise between linguistic and personal expression.

Initially, accounts of this genre offered a well-defended description of perception as a process of increasing abstraction (cf. Halle, 1985). The difference between phonetic form and canonical phonemic form decreed the initial conditions. Perception began with a sensory pattern, and the perceiver was obliged to transform it in order to resolve its phonemic attributes. The asynchronous distribution of acoustic correlates of a phoneme in a speech stream precludes a simple alignment of the sensory attributes and a canonical segmental series. In this model, several influences on the expressed form of speech must be undone before the segments can be exposed: the effects on the acoustic correlates of phoneme contrasts due to variation in the rate of production, the effects attributable to anatomical scale differences among talkers, the effects due to differential placement of emphasis, to variation in articulatory clarity, to foreign accent, and, of course, the effects due to co-production of sequential phonemes, syllables and words. In short, the characterization depicted a perceiver wielding stable standards – schemas – of the typical sensory presentation of the phonemes in the language, and applying a perceptual function to strip the instance-specific detail from an impinging sensory stream. Once a sensory sample was recast with sufficient abstractness, it was fit to match a stable linguistically-determined form.

Evidence from the listening lab had calibrated a perceiver's suppleness in adapting to the properties that drive the expressed form of speech to depart from a hypothetical abstract form. If some proposals relied on a dynamic that operated feature by feature

(Stevens, 1990), others described the comparison of segmental instances to prototypes (Samuel, 1982), and, in contrast to principles of likelihood, other accounts invoked a standard of segmental goodness independent of typicality yet still subject to the influence of experience (Iverson & Kuhl, 1995). The shared premise of these accounts is the use of progressive abstraction for the perceptual accommodation to variability. Categories of phonemic experience are rightly understood as commutable markers of contrast independent of talker or circumstance. After all, there is no pair of words that depends for its contrast on production by a specific talker, at a specific speech rate, paralinguistic expression of affect, vocal pitch, etc. But, the actual phonetic form of speech is too bound to the local conditions of production to be simply redeemed as an abstract phoneme series composing a word. The view that the incommensurate phonetic and phonemic forms are harmonized by reshaping the phonetic form into a less specific and more general version has been called *abstractionist* (Richardson-Klavehn & Bjork, 1988).

In such accounts, to appraise an unanalyzed bit of speech a perceiver must reconstruct the incoming sensory form to permit contact with a schematic idealization, or so an abstracting account could claim until critical studies of priming with spoken words. In a priming paradigm, the effect of a collateral probe (called "the prime") on the performance of a perceptual task is generally taken as evidence of relatedness. The closer the relation of a prime and a target, the greater the facilitation by the prime of a test subject's act concerning the target. This description of perception as the recognition of an abstract form warranted equivalence of the detailed phonetic variants of a spoken word used as prime and target because the point of contact inherent to identification was allegedly indifferent to the disparity among spoken instances of the same canonical phonemic form. But, in a series of studies that dislodged abstraction as the orthodox formula in speech perception, test subjects proved to be acutely sensitive to the exact phonetic similarity of prime and target, as if the specific phonetic attributes were preserved, and not simply registered as a preliminary to the process of abstraction requisite to identification (Goldinger, Luce, Pisoni, & Marcario, 1992; Luce, Goldinger, Auer, & Vitevitch, 2000).

In a description of perception by abstraction, the set of contact points is given by the number of resolvable types. The set is potentially small if the segmental phoneme inventory is used. If legal pairs or triads of phonetic segments are used, the set is larger, perhaps tens of thousands for English in comparison to the three dozen phoneme segments, but this set size can hardly be taxing on a nervous system capable of impressive feats of rote learning. But, imagine an indexing scheme representing instance-specific variation: it expands without limit. In contrast to the notion of the infinite use of finite means at the heart of every generative system, long-term knowledge that only encoded every raw instance is simply not compatible with the componential nature of phonology and morphology not to mention parity at any level. And, this consideration cannot apply solely to perceived form, for some studies had shown that we track the differential likelihood associated with the modality of the instances (Gaygen & Luce, 1998). That is, a spoken instance is encoded in a form distinct from a heard instance; a typed instance is marked in memory to distinguish it from a read instance. So far, there has been general agreement that these varied instances coalesce into types that match the abstract forms, preserving

the linguistic drivers of differentiation of lexical items through highly varied realization of canonical form. But, how are instances encoded?

In some descriptions of the adaptive resolution of superordinate phonemic types and subordinate phonetic instances, each level is treated as a linguistic representation derived from a raw sensory sample (Goldinger & Azuma, 2003). The instance is preserved as an unelaborated residue of stimulation. A literal understanding of an instance specific memory of utterances warrants a sensory encoding, for this is the only kind of representation that does not oblige the perceiver to an interpretation that substitutes for the direct experience of the instance. Yet, this notion can only be sustained in disregard of the psychoacoustic benchmarks of speech sounds (for instance, Pisoni & Tash, 1974). The unelaborated impression is gone in a tick of the clock. Indeed, the fleeting trace of an utterance arguably forces the retention of instance specific attributes while precluding an encoding of raw auditory experience.

We do not know the form of instance specific attributes yet, though some studies show that a perceiver is exquisitely sensitive to subtle phonetic variants, those that are far more detailed than simple categorization requires (McLennan, Luce, & Charles-Luce, 2003). Some phonetic variants are obviously due to chance–speech produced with food in the mouth, for instance, includes concurrent acts that compromise the expression of linguistic and paralinguistic properties with the accidental moment by moment acts to retain the bolus of food in the mouth. Other subtle phonetic variants are regulated, such as those that distinguish dialects and idiolects, and from their consistency we can infer that their production is perceptually monitored, and that phonetic perception incorporates dialectal and idiolectal dispositions, at least some of the time. Such sensitivity to varieties of phonetic expression at large in a language community might have played a role in findings that subphonemic discrimination of speech sounds always exceeded a prediction based on phoneme identification (see Liberman, 1957). Although these reports had been explained as an expression of auditory sensitivity, fine grain phonetic differences exist at a parallel level of resolution, and it is likely that perceivers attend to this detail because at this grain the linguistic and paralinguistic drivers of expression converge. A finding of instance specificity is potentially reducible to allophonic specificity, at least in linguistic dimensions of this phenomenon. But, not all specificity will be reducible to linguistically regulated properties of speech. In order to explain episodic properties of utterances – you were standing in the moonlight, the breeze was lightly rustling the leaves and a firefly twinkled just as you whispered, "Jazz and swing fans like fast music" – a state-dependent form of inscription might be required, but this is unlikely to be central to language. If this approach to speech perception holds potential for explaining the core problems that motivate research, perhaps because it is the most freewheeling of the accounts we have considered. Others lack the suppleness required by the accumulated evidence of the perception of speech as a cognitive function that finds linguistically specified contrasts under conditions that defy simple acoustic, articulatory, visual and tactile designation.

A listener who attends to subtle varieties of phonetic expression in speech is obliged to do so by the lack of uniformity in speech production. In accommodating this aspect of

variation, a listener meets a challenge created by language communities. The individuals who compose our communities vary in anatomical scale, dialect and idiolect, age, social role and attitude, and these dimensions are expressed in each utterance along with the linguistic message. If the sensory samples reflect these converging influences on expression, it is not surprising that a listener's attention to the attributes of a spoken event include features of the talker and the conditions in which an utterance occurred. In perceiving speech, a listener attends to personal attributes of the talker, and research on the perception of individuals, though it has sometimes run parallel to studies of linguistic perception, ultimately converges with it.

2. PERCEPTUAL IDENTIFICATION OF THE TALKER

Research on the perceptual organization of speech establishes that no set of auditory qualities is necessary for the resolution of linguistic form. A listener experiences linguistic impressions evoked by an indefinitely wide variety of acoustic causes. But, the quality of the voice, varying as widely as the acoustic causes of phonetic impressions, is undeniably salient. It is consistent with this intuition that some models have apportioned the perception of voice quality to a separate analyzer, one that is concerned with nonlinguistic attributes, including indexical properties of the talker producing the utterance. Such a clean separation of linguistic and indexical perceptual organization is a ready solution to the problem entailed in between-talker perturbations of linguistic form, the effects of which have been charted in numerous studies (see reviews in Johnson & Mullennix, 1997). A corollary question provoked by the separation of linguistic and indexical perception is whether auditory quality is necessary for the discrimination and identification of different talkers. Although the apportionment of functions warrants it, both forensic and laboratory investigations of talker identification and discrimination yield little hope of delineating a set of acoustic attributes or auditory qualities that designate individual talkers. Recent research on the effects of variation across individuals and instances in the perception of phonetic form indicate that indexical and linguistic processes are likely to be concurrent organizations of the same underlying cause.

The first empirical study of individual identification from speech noted the prevalence of voice recognition testimony by earwitnesses in court cases, despite a lack of scientific evidence attesting to a listener's ability to perform such a task (McGehee, 1937). Twenty years after Bell Laboratories created the sound spectrograph for speech analysis, Kersta (1962) developed a talker verification technique that involved visual analysis of spectrograms by trained experts, and coined the term *voiceprint* in hopeful analogy to the fingerprint. The contentious use of this technique in court cases and by government investigative agencies motivated much of the research on talker identification in the 1960s and 1970s (Hollien & Klepper, 1984; Kreiman, 1997); despite the controversy, earwitness evidence and expert analyst testimony are admissible in courts on a case-by-case basis – but not in Maryland. From the start, talker identification by ear was found to vary widely in accuracy. This reflects the fact that the acoustic products of an individual talker, while unique, are not distinctive. Arguably, the qualitative ways in which talkers differ

from each other and the way the speech of a single talker fluctuates in different situations derive from the same source: variability in phonetic repertoire. Without a dependable set of voice quality attributes to allocate to a separate stream, concurrent resolution of a word and a talker can be viewed as a form of multistability – obliging a listener's attentional finesse rather than segregation of analytical functions.

2.1. Talker Identification by Human and Machine

It seems effortless in ordinary circumstances to identify a talker by listening to speech, especially in a context of other commonplace events. Speech often occurs in situations in which conversants see one another, and even without visual contact, the uncertainty of a talker's identity can be constrained by other situational factors. For relatively long stretches of speech (>1200 ms) recorded under optimal conditions with a closed set of familiar alternative talkers, talker identification is nearly perfect, unsurprisingly (see reviews by Hollien & Klepper, 1984; Kreiman, 1997). Compromising any one of those ideals leads to a predictable decline in identification performance, and familiarity with the talker set and language are particularly important for accurate identification (Hollien, Majewski, & Doherty, 1982). Talker identification from speech can be difficult to explain because, unlike physical residues such as fingerprints, a talker's acoustic realization of the same word varies from instance to instance; and, no single acoustic attribute varies less across an individual's utterances than between individuals, precluding an account of identification by a simple acoustic feature. Talkers differ in multidimensional aspects of voice quality, and researchers have expended considerable effort attempting to find reliable psychoacoustic models of these differences (see Kreiman, Vanlancker-Sidtis, & Gerratt, 2005; Laver, 1980; Nolan, 1983). The partial success of this enterprise has yielded machine implementations of talker identification routines that are used in security applications.

The developers of automatic methods of talker identification might promise very high accuracy rates for their systems, similar to those found when human listeners recognize a familiar talker from speech. These methods benefit from the fact that a cooperative and motivated individual provides the standard and comparison samples, and the bank of standards is limited to a finite number of individuals. The main challenges for such systems are the possibility that an intruder could use a recording of the target individual to trick the recognizer, or that the circumstances during collection of a comparison sample deviate too far from those of the standard. For example, IBM's Voice Identification and Verification Agent boasts a 1% false alarm rate with only a 3% false rejection rate for a 20 s automated interview procedure (Navratil, Kleindienst, & Maes, 2000). Similar claims are made by the current industry leaders in voice identification technology, Nuance and SpeechWorks. The apparent success of these products arises as much from an individual's private knowledge as from a faithful acoustic rendering of the voice – these products rely on proprietary spectral analysis routines for acoustic voice verification, yet the system architecture employs an error-prone speech recognizer to verify answers to the interview questions. Although these systems can be found

in many corporate settings, companies have not completely forgone the use of live agents to handle false rejections from the routines, because automatic recognizers are so labile to subtle acoustic differences within the same talker. Despite claims to the contrary by the developers of the systems, the lack of more widespread application of such devices attests to the impracticality of their use.

2.1.1. Talker identification in forensic settings

In forensic applications of talker identification, an earwitness to a crime attempts to match a target talker from memory, or an expert analyst attempts to match a recorded sample of an unidentified talker to a set of known alternative talkers using a combination of auditory and visual spectrogram inspection. Under these conditions, it is impossible to calibrate accuracy because the set of alternative suspects might happen not to include the actual perpetrator. When a listener is not familiar with a talker, identification can be quite poor, especially with a potentially uncooperative suspect who might adopt a vocal disguise – although identification accuracy in laboratory settings is above chance across a set of listeners who are familiar with the talker (Hollien et al., 1982). Often, the target speech sample might not have been produced or recorded under quiet conditions, and the analyst must rely on degraded acoustic samples and subsequently poor spectrographic representations.

From its humble beginnings in the 1960s, there has been an effort to establish the admissibility of voiceprint analysis into court proceedings, with some success. At best, an earwitness or voiceprint examiner can choose the member of a voice line-up who sounds most similar to the original talker, or come to no decision. Expert voiceprint analysts often opt for no decision. However, for the victim of a crime providing earwitness testimony, the motivation to identify a perpetrator might override the uncertainty inherent in such a task, leading to a mistaken identification. Moreover, any decision based on relative similarity is extremely sensitive to the set of alternatives. To construct a fair test, a talker identification line-up should include a single target suspect paired against a set of talkers known to be innocent who are similar to the target in sex, gender, dialect, height, age, weight, socio-economic status and level of education. All of these factors have been found to form the basis for discrimination among talkers, and some are known to be identifiable with a moderate level of accuracy from speech alone (for a recent casting of this contentious issue, see Krauss, Freyberg, & Morsella, 2002; Morsella, Romero-Canyas, Halim, & Krauss, 2002).

2.1.2. Acoustic basis for talker identification

Do any specific acoustic parameters elicit a consistent impression of a talker's identity? Ideally, an acoustic attribute for talker differentiation would be a correlate of fixed anatomical differences between talkers, impervious to speaking habits and situational corruption. In order to identify an individual talker, not only must the parameter distinguish different talkers, but it must also be consistently produced by a given talker, or at least the variation in the production of this hypothetical attribute by a single talker should be smaller than

that which occurs between talkers. Previous reviews of talker identification evaluate the utility of many acoustic parameters, ranging from average phonatory frequency (Fø) to more derived measures of laryngeal and supralaryngeal acoustic effects (Hollien & Klepper, 1984; Kent & Chial, 2002; Laver, 1980; Nolan 1983). Most of these measures require about a minute of fluent speech to provide stable estimates of a single talker, the reliability of which could be due to the increase in sampled phonetic repertoire as well as the absolute amount of acoustic data (Bricker & Pruzansky, 1966).

Perhaps the most obvious differences between talkers are due to sex (male in contrast to female) and age (children in contrast to adults), which despite great overlap throughout the range are conspicuously expressed as differences in average Fø and formant frequencies, presumably due to differences in vocal anatomy. The large differences in average Fø among males, females and children (132 Hz, 223 Hz and 264 Hz; Peterson & Barney, 1952) have commonly been treated as differences in scale (see Fant, 1966). However, because an individual talker produces a wide range of variation in Fø that overlaps with other talkers' Fø distributions, a single measure alone is not enough to differentiate even males from females. Moreover, average Fø is not reliable over time for the same talker, so any difference that is large enough to differentiate talkers of the same sex is likely to typify the same talker at different times (Markel & Davis, 1979). An automatic routine that only exploited long-term properties of Fø would be vulnerable to the use of a different recording of the same talker at another time or under different circumstances, leading to a false rejection. Although including Fø aids perceptual identification relative to filtering or manipulating speech, it is not clear whether the improvement in identifying a talker is based on static long-term properties or dynamic short-term fluctuations of Fø (Nolan, 1983). Furthermore, because pitch is perceptually salient, it can be readily manipulated by a talker attempting to disguise the voice.

In supralaryngeal measures, female vowel formants are roughly 20% higher on average than male formants, and child formants are about 20% higher than adult female formants. However, the magnitude of these differences varies widely across both the individual formants (F1 < F2 or F3), the vowel inventory ([ɔ]<[æ]), and different languages, therefore, they can not be a simple function of talker anatomy (Fant, 1966; Johnson, 2005). There is no characteristic formant profile for an individual talker that is stable over time. In order to eliminate the sample bias of variation caused by different assortments of individual segments on the measurement of formant frequencies, long-term spectra have been studied. These aim to characterize a talker independent of phonetic samples by averaging the short-term power spectrum over a long span of speech, generally at least 10 s of continuous speech (Furui, Itakura, & Sato, 1972). Like average Fø and formant frequency estimates, long-term spectra are not reliable for indexing an individual talker over different speaking contexts. Compared to identification by human listeners hearing familiar talkers, a talker identification routine based on long-term spectra is more severely affected by distress and disguise (Hollien et al., 1982). Long-term spectra do reflect differences in voice quality, but as correlates of individual talkers they are not stable, failing to track the same talker under different speaking conditions. Figure 3 illustrates long-term spectral measures adapted from

Nolan (1983). For example, one talker adopting two different supralaryngeal postures, neutral or pharyngealized, shows as much change between postures as another talker adopting the same contrast; and, the global change is not uniformly reflected in the long-term spectral consequences for each talker. As shown in the bottom plot, additional derived measures of long-term spectra fail to distinguish an individual talker consistently across different contexts of speaking – for some articulatory postures, the two talkers overlap in the log-transformed slope of the long-term spectra, while for other postures, the talkers are very distinctive.

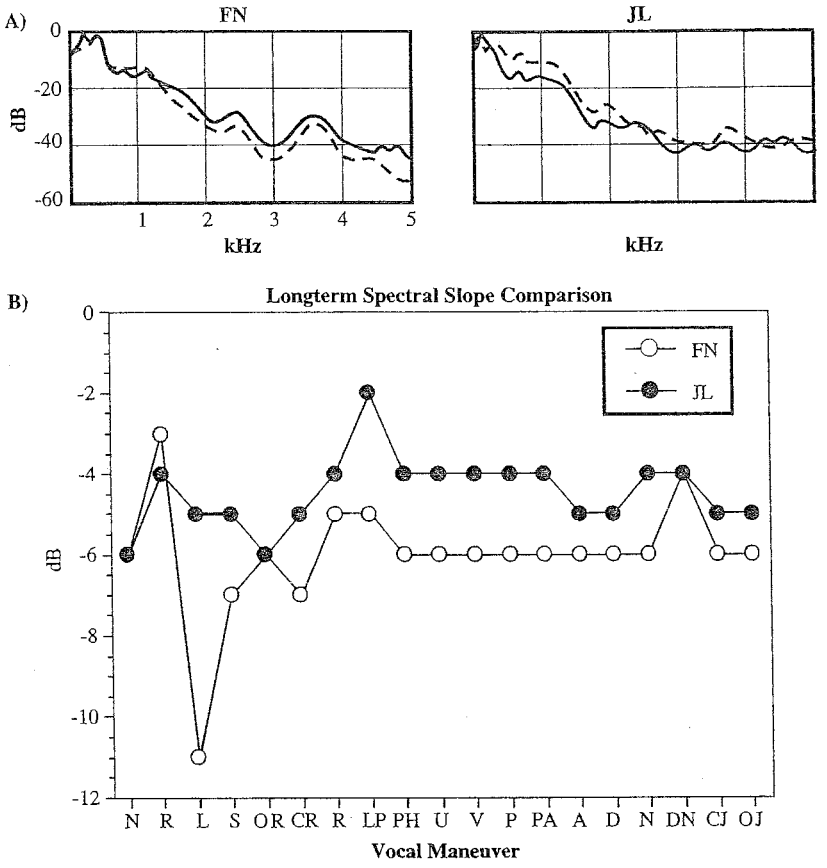

Figure 3. The effect of selected vocal postures on long-term speech spectra (Nolan, 1983). (A) Neutral and pharyngealized spectra for talkers FN and JL; (B) Effect of vocal maneuvers on the slope of the speech spectrum for talker FN (open bullet) and JL (filled bullet): n=neutral; r=raised larynx; l=lowered larynx; s=labial/spread lips; or=labial/open rounding; cd=labial/close rounding; r=retroflex; lp=laryngo-pharyngealized; ph=pharyngealized; u=uvularized; v=velarized; p=palatalized; pa=palato-alveolarized; a=alveolarized; d=dentalized; n=nasalized; dn=denasalized; cj=close jaw; and oj=open jaw.

2.1.3. Phonetic identification of talkers

The differences in long-term spectral measures of an individual talker across different postures are roughly equivalent to those between talkers. Despite an impression that such complex acoustic measures provide a context-free representation of a talker's unique vocal profile, they are not characteristic of the talker because they do not designate the talker independently of different vocal postures. Furthermore, many qualitative supralaryngeal vocal postures involve anatomical structures used in phonetic expression, influencing the production of phonetic forms. Laver (1980) argues that it might be possible to derive the postures from patterns of variability in phonetic expression, making voice quality a function of phonetic form.

Because of the versatility of the speech production system, Nolan suggests that voice quality measures may best be used to classify rather than to identify talkers absolutely. Indeed, many empirical projects point out the lack of reliability in judgments of talker identity based on impressions of voice quality (see review in Kreiman, 1997). Voice quality is not an independent perceptual stream, divorced from phonetic attributes, that can be used to identify the speaker of a message. Even nasal resonances, long considered good cues for talker identification due to the relative anatomical immutability of the nasal cavities (Glenn & Kleiner, 1968; Sambur, 1975; Su, Li, & Fu, 1974), are susceptible to many situational factors, including everything from colds to stylistic variation.

Some empirical studies of perception provided exclusively dynamic acoustic attributes of speech, using a synthesis technique to exclude a contribution of Fø and familiar voice quality to perception. Under such extreme conditions, listeners still resolved phonetic attributes and talker identity (Remez et al. 1997). In these studies, the three lowest formants of speech were replaced with time-varying sinewaves, with a fourth tone following the fricative poles. These signals lack the natural acoustic correlates of vocal sound production, yet the linguistic message was readily resolved; the acoustic transformation to sinewave tones evoked fine grain phonetic properties. Listeners could identify both familiar and unfamiliar talkers in closed set sinewave tests, and the performance by listeners who knew the talkers was similar to those who simply matched a natural sample to a sinewave utterance. Arguably, a listener who recognized a sinewave version of a familiar talker must have relied on longstanding knowledge of a particular talker's articulatory habits, or idiolect, which comprise linguistic, not qualitative, attributes. It appears that a listener will use any parameter, even a phonetic one, that can identify a talker. No acoustic attribute is necessary, and each parameter varies in its sufficiency for contributing to the distinction across different situations.

As with identification of linguistic attributes of speech, talker identification is not reliably cued by a particular set of acoustic attributes. Instead, listeners are evidently not particular about the acoustic attributes or the psychoacoustic effects that are used to identify specific individual talkers. Neither the pitch or quality of the voice, nor the short- or long-term spectrum, nor any specific set of acoustic correlates of a talker is required for identification. Apparently, a listener identifies attributes of a talker as if the commitment

to the particular sensory manifestation of the individual is adaptable. In like manner to the perception of phonetic attributes, this inclination to find functional contrasts in unexpected acoustic or auditory form opposes the fixity of a normative conceptualization of talker identification.

2.2. Perception of Words and Talkers

A significant portion of the literature on talker differences has considered this source of variation a kind of noise in the correspondence between phonetic and acoustic attributes introduced during the transformation from phoneme to phonetic event. From this perspective, talker differences obscure the acoustic correlates of phonemes that permit a perceiver to resolve words. Many accounts of speech perception propose a normalization function that removes this noise from the signal, arriving at abstract phonemes (Pisoni, 1997; Magnuson & Nusbaum, 2004; Johnson, 2005). This presumes that a listener does not use phonetic attributes to identify a talker, which are stripped from the sensory properties of speech by such functions. A phoneme representation of speech lacks all trace of the circumstances of its uttering, including those phonetic attributes that are distinctive of dialect and idiolect. The abstraction obliged in such accounts leaves only voice quality available for talker identification. However, not only is voice quality an unreliable metric for describing or identifying talkers, it is not necessary for identification of talkers, as shown in studies using sinewave speech. Laver (1980) proposed a set of physiological parameters that can be used to distinguish voice qualities, and many of these overlap with the gestural/articulatory systems used to differentiate phonemes as well (Goldstein & Fowler, 2003; Studdert-Kennedy & Goldstein, 2003). Indeed, a talker's idiosyncratic phonetic repertoire converges on the same series that a listener apprehends as a word.

Preliminary attempts to characterize talker differences aimed to quantify the effect of a particular vocal tract on production of a phone class. Gauging the large acoustic phonetic differences between men, women and children was a good place to start, but such inventories failed to yield a uniform function attributable to vocal tract size (Fant, 1966; Johnson, 2005). In order to remove the effect on a phoneme or word of the sex of the talker, for example, a different scaling function would be necessary for words containing a rounded vowel like /u/ than for an unrounded vowel like /ɑ/, and for individual formants themselves. A hypothetical scaling function that varies with the phonemic class of the segment is of limited value, and consequently it appears unlikely that perceptual rescaling with the sex of a talker plays a role in identifying words, if at all. Acknowledging that talkers also differ in habitual vocal tract postures, such as larynx lowering or raising, retroflexion, or velarization, presents a similar barrier to deriving a talker normalization function. Once again, phonemes differ in their susceptibility to such postures, roughly depending on the relative difference between the anatomical focus of the phonemic and qualitative attributes (Laver, 1980). Perception of words and talkers alike might depend on resolution of detailed phonetic form, which can evoke both phonological and idiosyncratic indexical impressions.

Some recent empirical projects demonstrate that these idiosyncrasies are not always discarded early in the projection of sensory samples to phonemic impressions – they

survive transduction to affect perception of the message. To consider several examples, both implicit and explicit memory of spoken words in a list are affected by whether a single talker or several talkers utter the items (Goldinger, 1996; Nygaard, Sommers, & Pisoni, 1992). Familiarity with a talker differentially facilitates resolution of the words spoken by the talker independent of likelihood or familiarity (Clarke & Garrett, 2004; Goldinger, 1998; Nygaard, Sommers, & Pisoni, 1994; cf. Lieberman, 1963). Failure to notice a change in the talker during a shadowing task has no effect on speed of shadowing, while noticing the change does (Vitevich, 2003). Listeners track differences in voicing, exhibiting graded rather than uniform perceptual categorization of consonants, and the category structures converge on talker differences (Allen & Miller, 2004; Miller & Volaitis, 1989). A listener will even track subcategorical variation in the precise timing of voicing contrasts, and reciprocate them by approximating the temporal expression of voicing in shadowed responses (Fowler, Brown, Sabadini, & Weihing, 2003).

A talker is characterized by more than a collection of simple acoustic attributes imposed upon a word or phoneme, whether such attributes derive from anatomy or from habit, and listener sensitivity to these effects does not take the form of a normalization function yielding only linguistic categories. Talker conditioned within-category variation in phonetic form influences speech perception and memory, at the same time that variants in phonetic form index the talker.

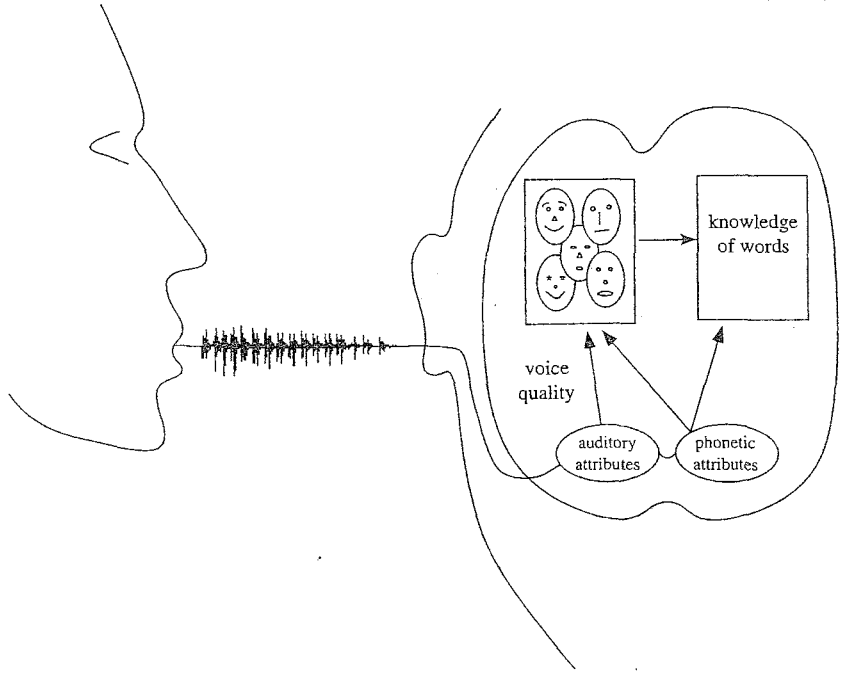

Figure 4. Hypothetical system architecture for the perception of individual talkers from qualitative and phonetic attributes (Remez, Fellowes, & Rubin, 1996).

2.3. Individual Stylistic Variation

The attempt to delineate the acoustic attributes correlated with anatomical differences was reasonable because anatomical differences among talkers were thought to be a kind of scalar property limiting the range of other factors that might influence speech production. Changes in a talker's anatomy surely occur – due to growth and maturation, tooth gain or loss, dueling injury – but, roughly, we are stable over long spans. If the dimensions of anatomy are relatively secure, why does the speech of an individual vary across different contexts? Notwithstanding the effects of the compliance of the pharyngeal cavity and the cheeks, tongue and lips on acoustics, much of the moment-to-moment variability that an individual talker exhibits also reflects changing goals, situations and addressees. A talker might produce more or less characteristic attributes of a regional dialect on different occasions (Bourhis & Giles, 1977; Giles, 1973; Labov, 1966, 1986). A talker's emotional states are conveyed in speech, although the acoustic effects are exceedingly complex (Banse & Scherer, 1996; Scherer, 1986; Williams & Stevens, 1972). Talkers shorten the duration of repeated referents in discourse, and listeners attribute such shortening to *givenness*, that is, to a topic already introduced in a narrative (Fowler & Housum, 1987). A large component of research on within-talker differences attempts to characterize the acoustic changes under conditions of clear versus casual speech production (Picheny, Durlach, & Braida, 1985).

The ability to produce clear speech under appropriate conditions demonstrates a talker's ability to control detailed aspects of phonetic form. As surveyed by Uchanski (2005), the main acoustic-phonetic changes from conversational to clear speech include increases in amplitude, duration and Fø; changes in the slope of the long-term spectrum; phonological effects such as decreased vowel reduction, more prominent stop burst production and elimination of alveolar flapping in coronal consonants; and increases in F1 and F2 frequency and exaggerated voicelessness. Once again, the phonetic differences between the casual and clear speech of an individual deploy the same set of parameters that distinguish different talkers. Of course, when a talker speaks clearly in order to indicate sincerity, or authority, or to reduce the hazard of misunderstanding, this ultimately finds expression in an individual talker's idiolectal phonetic repertoire. And, clear speech for one talker might not be clear for another (Bradlow & Bent, 2003).

Taken together, these phenomena demonstrate the inherent flexibility of language – listeners readily adapt to changing talkers, while still perceiving differences among them. Although comprehension requires relative stability in linguistic form, not all variability must be discarded in service of parity. A listener must track the talker's use of linguistic currency in order to calibrate the tokens available for exchange and to determine what the talker means from what the talker says. With a detailed resolution of phonetic form, speech perception buys both linguistic types and indexical properties. Attention plays this role in speech perception, whether of linguistic or indexical attributes, and the next section describes a self-regulatory process that sets the grain of speech perception with an ear toward speech production.

3. PERCEPTUAL SELF-REGULATION

In dexterity, complexity and effectiveness, there is little that we do that surpasses the production of speech. This is not to deny that each of us is occasionally tongue-tied, incomprehensible or wrong. But, apart from the challenge of imagining something useful or clever to say, articulation is normally fast and easy, and on this dimension the role of perception in production was initially misunderstood. The origin is in Lashley's analysis of acts that express a sequential order, described in a founding document of psycholinguistics (Lashley, 1951). He had turned his attention to language because of the empirical opening it offered to oppose the prevailing models of action, which evaded the problem of serial order in a conceptualization of coordinated movement restricted to peripheral sensory-motor chains. Among the arguments in his analysis entailing plans for sequential acts, Lashley contemplated the fluency with which the articulatory components of speech succeed each other. At the phonemic grain, the procession of expressed elements occurs so rapidly that the sensory consequences of the first cannot be responsible for triggering the production of the next, and so on. The second component in a series is initiated, in Lashley's line of reasoning, well in advance of the conduction of the afferent consequences of the preceding component, whether taken as auditory, orofacial tactile or muscle sense. In this circumstance, he proposed that a series of acts was composed and controlled centrifugally, without slowing to the crawl obliged by a sequencing mechanism of sensory triggering. He was right in part.

Phonetic expression is incredibly stable. Over decades, habits of articulation are consistent (House & Stevens, 1999), and this stability reflects the constancy of anatomical and functional constraints on vocal acts. That is to say, skeletal actions must be adaptable to an environment of displaceable masses and extrinsic forces. In contrast, the topography of vocal landmarks – teeth, palate, tongue – is relatively fixed, and the masses intrinsic to the vocal tract vary only gradually over the lifespan and not at all during an act. If the microenvironment composed by a vocal tract is durable, the vocal acts committed within it need not be especially adaptable. From this perspective, it is not surprising that research pursuing Lashley's speculation found ample evidence that speech production is hampered very little by the absence of sensation.

The research corroborating this viewpoint about articulatory control aimed to assess the consequences of disrupted sensation on speech, and the findings were consistent. To interfere with auditory experience of speech, a noise load is imposed, masking a listener's self-productions (Lane, Catania, & Stevens, 1961); to interfere with orofacial taction, the lingual, labial and pharyngeal surfaces are anesthetized (Borden, Harris, & Oliver, 1973); to interfere with muscle sense, the intrafusal muscle fibers supplied by the mandibular branch of the trigeminal nerve are selectively albeit reversibly blockaded (Abbs, 1973). None of these sensory streams is essential to the control of speech production. It is easy to see that such studies would warrant an account of the articulation of speech in central open-loop control, given this juxtaposition of prodigious expression in both variety and fluency, and relative proficiency of articulation in conditions that preclude sensory monitoring.

If an adult can succeed in producing speech without sensory supply, this does not mean that language development can occur without the means to align self-produced vocalization with the speech of others, whether in the early home setting of infancy or in the schoolyard society of juveniles. After all, though the internal environment of the vocal tract is set, a conversational environment is often distressingly unpredictable. The deterioration of spoken language that can follow deafening in childhood (Binnie, Daniloff, & Buckingham, 1982) is evidence of the importance of auditory reafference; in the classic contrast, the causes of sensory states are dichotomized as *reafferent,* owing to the consequences of self-produced action, and *exafferent*, owing to extrinsic objects and events independent of self-produced action (von Holst & Mittelstaedt, 1950). Aside from developmental functions in which the exafferent effects of the speech of a linguistic community are reprised in the reafference of a young talker, there is ample evidence that self-regulation of phonetic production depends on reafference which, though inessential in the short-term, is exploited nonetheless when it is available. This theme in speech perception research is evolving, though it is possible to see the principles emerging in studies of detailed phonetic production. Throughout, the mark of reafferent control of speech is the constraining effect of linguistic repertoire on sensitivity and production alike.

3.1. Vocal Effort

The original report by Étienne Lombard (1911) of reafferent control of speech production pertained to vocal effort. A talker adjusts the power of speech as if to maintain a constant difference between the sensory effects of the voice and the momentary extrinsic noisemakers obscuring conversation. Reviewing studies of this *Lombard sign*, Lane and Tranel (1971) reported that in addition to reafferent control, the functions exhibited an intriguing communicative contingency, namely, compensation in production for noise load occurred only when a listener was present. If compensation were purely an egocentric self-regulatory function, it would appear regardless of the presence of an audience. However, talkers did not regulate vocal power when reading a script into a microphone with noise presented over headphones; only conversational settings induced the critical adaptive pattern. Moreover, vocal effort is also regulated by a crossing factor, the perceived distance between the talker and listener (Liénard & Di Benedetto, 1999). The standards for regulating appropriate vocal effort involve power in relation to noise corrected by the talker's implicit goal of maintaining the sound level at the listener's ear. Some portion of this regulatory ability is not restricted to humans producing speech: Zebra finches also adapt vocal effort to ambient noise (Cynx, Lewis, Tavel, & Tse, 1998).

3.2. Phonation

Reafferent control of vocal effort is a relatively gross if complex parameter, and not the only evidence of reafferent regulation. Studies of more detailed control indicate that the pitch of the voice falls under reafferent control. In one line of research on this aspect of production (Jones & Munhall, 2000), a subject articulated a long syllable under

conditions in which the acoustic experience of self-produced Fø provided by the experimenters was veridical or altered electronically. When phonatory frequency was modified, it was transposed up or down. Participants in the study compensated for the displacement, opposing the perturbations established electronically. For instance, when pitch was shifted upward in frequency, a talker lowered vocal pitch to bring reafferent experience to the internal standard.

3.3. Phonetic Production

Critical findings about the regulation of fine segmental structure are also reported. In one, Houde and Jordan (1998) created an acoustic-phonetic analogy to studies of visual perceptual adaptation using displacing prism spectacles. In the visual circumstance, viewers gradually recalibrate their reaching movements to accommodate the illusory displacement of the visible world created by the prisms. Moreover, immediately after removing the prisms, reaching shows an adaptive rebound of the discontinued visual displacement. In auditory displacement during speech production, Houde and Jordan provided listeners with online acoustic perturbations of their own utterances. Over the course of a prompted, whispered syllable production task, talkers heard the natural acoustic consequences of vocal sound production, but the speech spectrum was modified electronically; the method induced a tolerable 16 ms delay. The modification entailed shifting formant frequency so that a talker heard a different vowel in response to the produced vowel: /ɛ/ formants were shifted either up to /i/ or down to /æ/, always produced in a /bVb/ syllable. In the context of other consonants, /ɛ/ formants were unaltered, and formants from a talker's production of other vowels were likewise unaltered. While receiving perturbed feedback, most talkers shifted their productions of the altered vowels to compensate for the distortion, resulting in production of lowered or raised vowels, as appropriate to counter the formant frequency displacement. These altered productions, once shifted in frequency by the electronic apparatus, matched the reafference typical of the intended vowel produced in the clear. Furthermore, compensation generalized to production of the same vowel in different consonant contexts from the training set and to different vowels, and talkers persisted in shifted productions when the altered reafference was replaced with masking noise.

Reafferent control is not limited to auditory samples, an aspect of the multimodal sensory expression of phonetic attributes. In one instance of the effectiveness of tactile and muscle sense, a talker was outfitted with a dental prosthesis effectively lengthening the maxillary incisors (Jones & Munhall, 2003). A variety of changes in phonetic expression would be expected following a derangement of familiar dimensions, of course, and this study focused on instances of /t/. A change in a stable feature of the vocal anatomy can be expected to disrupt articulation, and to impose a requirement to adapt specific aspects of sound production. A subject in the study improved in producing natural-sounding /t/ as a result of experience with the dentures over a testing hour when auditory reafference was blocked; once auditory reafference was allowed, there was little benefit beyond that which somatic reafference established. In its time course, the slow adaptation to an abrupt

change in a fixed vocal structure is familiar from our own experience as dental patients. It can take a while to get used to speaking with a bite splint or new choppers, and auditory reafference of the altered production that ensues is motivating but inadequate to elicit immediate compensation.

Adaptation of speech production to the presence of a pseudopalate that changes the shape of the roof of the mouth exhibits a similar time course and impact on fine placement (Baum & McFarland, 1997; McFarland, Baum, & Chabot, 1996). In contrast, a functional constraint, such as occurs when the motion of the jaw is fixed during speech production by the concurrent requirement to hold an object between the teeth, is readily assimilated. Although some individuals fare better than others in adapting to functional limits, in general reafference rapidly drives production toward typical phonetic expression. Indeed, although canonical form might be difficult to realize with the jaw height or the lip aperture fixed and motionless, a talker is not inexperienced in reconciling the functions of ingestion, deglutition and respiration with sound production, and either practice or endowment are exploitable to minimize the effect of this sort of limit in action.

Of the studies examining the effect of altered auditory reafference, the introduction of systematic departures from veridical samples provided evidence about the contribution of sensory states to phonetic production. The alterations created by researchers are subtle, though, in comparison to the drastic departure from veridicality experienced by a user of a cochlear implant. An implant user typically becomes accustomed to the anomalous quality delivered by the stimulator, an electrode that uses a coarse grain place-code to evoke a rough correlate of incident spectrotemporal acoustic properties in the activity of the auditory afferents. Pitch experience is evoked only poorly if at all, and experience of melody and harmony is meager in contrast to meter and rhythm. Nevertheless, adults who rely on such recurrent sensory qualities to control speech production can perform well (Vick et al., 2001). After a year of use the regulation of production extends throughout the English phone classes, and despite variation in success with adventitiously deafened linguistically competent adults the preponderant outcome observed in one survey establishes the value of reafference in sharpening phonetic production (Gould et al., 2001). Intuitively, an adult who already expresses and comprehends language might seem to have an advantage of experience in exploiting an impoverished sensory sample of speech, whether to perceive the speech of others or to produce speech by reafferent regulation. But, astonishingly, some children whose deafness is profound by age 3 can employ the reafference available through an implant to speak and to comprehend speech well enough to master English phoneme contrasts (Svirsky, Robbins, Kirk, Pisoni, & Miyamoto, 2000). It is unclear why so many children fail, though the success of many who can learn language this way defines the problem as linguistic rather than one specifically of sensory veridicality (Watson, Qiu, Chamberlain, & Li, 1996).

3.4. Self-Regulation in Conversation

One prominent claim about the production of speech in conversational conditions holds that production is regulated allocentrically as well as egocentrically. That is to say,

the talker regulates production with respect to the listener's state, and not simply with regard to the reafference associated with self-produced speech. In the elaboration by Lindblom (1990), conversation is described as a competition, in which a talker aims to minimize the effort in articulation by neutralizing contrasts under conditions in which the cost of miscommunication is offset by a listener's vigilance and acuity. The listener, complementarily, aims to minimize the cognitive load of vigilance and acuity by relying on the talker to produce salient versions of the linguistic contrasts. This talking version of prisoner's dilemma characterizes the self-regulatory options for conversations, and views the participants as not entirely cooperative, to its great merit. But, it does discount the evident sharing that occurs when people talk.

Studies of interacting talkers have found fairly consistent patterns of linguistic change over the course of conversational interaction, and such changes are variously termed coordination (Clark, 1996), alignment (Pickering & Garrod, 2004) or accommodation (Shepard, Giles, & Le Poire, 2001). Most of these projects examined the increase in similarity (convergence) of diverse aspects of interlocutor's speech, from the schematic (Garrod & Doherty, 1994), to the syntactic (Branigan, Pickering, & Cleland, 2000), to the lexical/semantic levels (Krauss & Weinheimer, 1964; Wilkes-Gibbs & Clark, 1992). Research on convergence at sub-lexical levels includes measures of acoustic-phonetic attributes such as perceived accentedness, fundamental frequency covariation, and voice amplitude (Giles, 1973; Gregory, 1990; Natale, 1975). Convergence in such parameters appears to be particularly susceptible to the effects of social factors that are confined to communication exchanges, such as interlocutors' relative dominance or perceived prestige (Gregory, Dagan, & Webster, 1997; Gregory & Webster, 1996).

Some reports also describe a pattern opposite to convergence under some circumstances (Giles, Coupland, & Coupland, 1991; Shepard et al., 2001). Although it is tempting to attribute convergence to an automatic imitative function facilitating increased intelligibility (e.g., Pickering & Garrod, 2004), divergence often does not preclude intelligibility, but serves a communicative need for the diverging party (Bourhis & Giles, 1977; Labov, 1974). Indeed, a recent study found that interacting talkers differed in the extent to which they converged in phonetic repertoire over the course of a single cooperative task (Pardo, 2006). The study employed a modified version of the Map Task (Anderson et al., 1991), in which paired talkers are given matched sets of iconic maps with labeled landmarks; in one set, the maps contain paths around various landmarks, while the other set of maps have only labeled landmarks. The talkers cannot see each other or their partner's maps, yet they must communicate effectively enough that the instruction receiver can duplicate the paths on the instruction giver's maps. The talkers in this study were able to perform the task well, and they also converged in phonetic repertoire over the course of the short interaction, but instruction givers converged to instruction receivers more than the reverse. Furthermore, the shifts in phonetic repertoire persisted to a recording session conducted immediately after the conversational setting. Because some talkers converged more than others, phonetic assimilation cannot be attributed to an automatic function in which perception of an addressee's phonetic variants primes recurrent production of the variants. Rather, to account for the asymmetry,

consider that perception yields phonetic variants available for selection, depending on paralinguistic functions operating in parallel with linguistic functions. In this case, those functions related to the talker's role in the conversational setting. Given that the phonetically variable lexical tokens were perfectly intelligible to both parties, it is arguable that phonetic convergence serves communicative purposes beyond that required for intelligibility – a listener tolerates a wide variety of acoustic–phonetic forms when perceiving speech produced in conversational interaction.

The inherent tolerance of variability licenses each talker to sample a broad contrast space to compose a unique phonetic repertoire. Conceivably, once an idiolect forms, some parameters remain relatively free to assimilate in social exchange. An elaboration of the circumstances that promote or preclude such compliance would provide a clearer view of the processes shaping the phonetic landscape from its linguistic and paralinguistic bases. Overall, the consideration of perceptual self-regulation in production exposes the accommodating relationship of perception and production despite the difference in grain. This asymmetry of range and detail – because we can do so little in comparison to what we discern – precludes a rigid coupling of perceived and produced speech, mirror neurons notwithstanding (Rizzolati & Craghero, 2004). Instead, attention at the phonetic grain informs both linguistic and paralinguistic functions, evoking linguistic forms that are regulated to fulfill the aims of communicative exchanges.

4. A CONCLUDING WORD

In this discussion of the theoretical and empirical study of the perception of speech, we have emphasized the principles motivating classic and recent explanations. Instead of a comprehensive review of research paradigms within this lively domain, we have adopted a functional focus on speech as an expression of language. As a consequence of this commitment, speech perception is cast as the means by which spoken expressions are resolved as coherent audible and visible events, as strings of familiar linguistic forms, as the product of an individual talker, and as a model for recurrent self-expression. Ostensibly, these related aspects of speech perception are the ordinary experience of talkers and listeners, and do not depend for their justification on specific research methods or puzzles. New investigations will determine whether the perspective articulated here proves to be useful in the search for new evidence to refine the descriptive and explanatory inquiry into the perception of speech.

ACKNOWLEDGEMENT

We are grateful for the generosity of our colleagues whose instruction and advice were helpful in setting the focus of our discussion: Carol Fowler, Louis Goldstein, Michael Studdert-Kennedy, David Pisoni, and Philip Rubin. We also offer special thanks to Daria Ferro for her steady hand and clear eye in drafting the artwork. This research was supported by a grant from the National Institute on Deafness and Other Communication Disorders (DC00308) and a grant from the National Science Foundation (0545133) to Barnard College.

REFERENCES

Abbs, J. (1973). The influence of the gamma motor system on jaw movements during speech: A theoretical framework and some preliminary observations. *Journal of Speech and Hearing Research, 16*, 175–200.

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America, 115*, 3171–3183.

Aman, R. (1980). Clean up your fexing language! Or, how to swear violently without offending anyone. *Maledicta, 4*, 5–14.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., & Weinert, R. (1991). The HCRC Map Task corpus. *Language and speech, 34*, 351–366.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*, 614–636.

Baum, S. R., & McFarland, D. H. (1997). The development of speech adaptation to an: artificial palate. *Journal of the Acoustical Society of America, 102*, 2353–2359.

Bernstein, L. E., Auer, E. T., Jr., & Moore, J. K. (2004). Audiovisual speech binding: Convergence or association. In: G. A. Calvert, C. Spence, & B. E. Stein (Eds), *Handbook of multisensory processing* (pp. 203–223). Cambridge, MA: MIT Press.

Bertelson, P., Vroomen, J., & de Gelder, B. (1997). Auditory-visual interaction in voice localization and in bimodal speech recognition: The effects of desynchronization. In: C. Benoît, & R. Campbell (Eds), *Proceedings of the workshop on audio-visual speech processing: Cognitive and computational approaches*, Rhodes, Greece, ESCA (pp. 97–100).

Binnie, C. A., Daniloff, R. G., & Buckingham, H. W. (1982). Phonetic disintegration in a five-year-old following sudden hearing loss. *Journal of Speech and Hearing Disorders, 47*, 181–189.

Borden, G. J., Harris, K. S., & Oliver, W. (1973). Oral feedback, I. Variability of the effect of nerve-block anesthesia upon speech. *Journal of Phonetics, 1*, 289–295.

Bourhis, R. Y., & Giles, H. (1977). The language of intergroup distinctiveness. In: H. Giles (Ed.), *Language, ethnicity and intergroup relations* (pp. 119–135). London: Academic Press.

Bradlow, A. R., & Bent, T. (2003). The clear speech effect for non-native listeners. *Journal of the Acoustical Society of America, 112*, 272–284.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition, 75*, B13–B25.

Bregman, A. S. (1990). *Auditory scene analysis*. Cambridge, MA: MIT Press.

Bricker, P., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *Journal of the Acoustical Society of America, 40*, 1441–1450.

Browman, C., & Goldstein, L. (1991). Gestural structures: Distinctiveness, phonological processes and historical change. In: I. G. Mattingly, & M. Studdert-Kennedy (Eds), *Modularity and the motor theory of speech perception: Proceedings of a conference to honor Alvin M. Liberman* (pp. 313–338). Hillsdale, NJ: Lawrence Erlbaum Associates.

Carlyon, R. P., Cusack, R., Foxton, J. M., & Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation, *Journal of Experimental Psychology: Human Perception and Performance, 27*, 115–127.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America, 25*, 975–979.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America, 116*, 3647–3658.

Cynx, J., Lewis, R., Tavel, B., & Tse, H. (1998). Amplitude regulation of vocalizations in noise by a songbird, *Taeniopygia guttata. Animal Behaviour, 56*, 107–113.

Diehl, R. L., Kluender, K. R., Walsh, M. A., & Parker, E. M. (1991). Auditory enhancement in speech perception and phonology. In: R. R. Hoffman, & D. S. Palermo (Eds), *Cognition and the symbolic processes: Applied and ecological perspectives* (pp. 59–76). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Eimas, P. D., & Miller, J. L. (1992). Organization in the perception of speech by young infants. *Psychological Science, 3*, 340–345.

Elliott, L. L. (1962). Backward and forward masking of probe tones of different frequencies. *Journal of the Acoustical Society of America, 34*, 1116–1117.

Fant, C. G. M. (1962). Descriptive analysis of the acoustic aspects of speech. *Logos, 5*, 3–17.

Fant, G. (1966). A note on vocal tract size factors and nonuniform F-pattern scalings. *Speech Transmission Laboratory Quarterly Progress and Status Report, 4*, 22–30.

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory & Language, 49*, 396–413.

Fowler, C. A., & Housum, J. (1987). Talkers' signaling of 'new' and 'old' words in speech and listeners' perception and use of the distinction. *Journal of Memory & Language, 26*, 489–504.

Furui, S., Itakura, F., & Sato, S. (1972). Talker recognition by longtime averaged speech spectrum. *Electronics and Communications in Japan, 55-A* (10), 54–61.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (in press). The Motor theory of speech perception reviewed. *Psychonomic Bulletin and Review*, *00*, 000–000.

Garrod, S., & Doherty, G. (1994). Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*, *53*, 181–215.

Gaygen, D. E., & Luce, P. A. (1998). Effects of modality on subjective frequency estimates and processing of spoken and printed words. *Perceptions & Psychophysics*, *60*, 465–483.

Gedda, L., Bianchi, A., & Bianchi-Neroni, L. (1955). La voce dei gemelli: I. Prova di identificazione intrageminale della voce in 104 coppie (58 MZ e 46 DZ). [The voice of twins: A test of intratwin identification of the voice in 104 pairs (58 MZ and 46 DZ).] *Acta Geneticae Medicae et Gemellologieae*, *4*, 121–130.

Giles, H. (1973). Accent mobility: A model and some data. *Anthropological Linguistics*, *15*, 87–109.

Giles, H., Coupland, J., & Coupland, N. (1991). *Contexts of accommodation: Developments in applied sociolinguistics*. London: Cambridge University Press.

Giles, H., Scherer, K. R., & Taylor, D. M. (1979). Speech markers in social interaction. In: K. R. Scherer, & H. Giles (Eds), *Social markers in speech* (pp. 343–375). Cambridge: Cambridge University Press.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1166–1183.

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

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics*, *31*, 1–16.

Goldinger, S. D., Luce, P. A., Pisoni, D. B., & Marcario, J. K. (1992). Form-based priming in spoken word recognition: The roles of competition and bias. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *18*, 1211–1238.

Goldstein, L., & Fowler, C. A. (2003). Articulatory phonology: A phonology for public language use, In: A. S. Meyer, & N. O. Schiller (Eds), *Phonetics and phonology in language comprehension and production: Differences and similarities* (pp. 159–207). Berlin: Mouton de Gruyter.

Gould, J., Lane, H., Vick, J. C., Perkell, J. S., Matthies, M. L., & Zandipour, M. (2001). Changes in speech intelligibility of postlingually deaf adults after cochlear implantation. *Ear & Hearing*, *22*, 453–460.

Gregory, D., Dagan, K., & Webster, S. (1997). Evaluating the relation of vocal accommodation in conversational partners' fundamental frequencies to perceptions of communication quality. *Journal of Nonverbal Behavior*, 21, 23–43.

Gregory, D., & Webster, S. (1996). A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status predictions. *Journal of Personality & Social Psychology, 70,* 1231–1240.

Gregory, S. W. (1990). Analysis of fundamental frequency reveals covariation in interview partners' speech. *Journal of Nonverbal Behavior, 14,* 237–251.

Grossberg, S. (2003). Resonant neural dynamics of speech perception. *Journal of Phonetics, 31,* 423–445.

Halle, M. (1985). Speculations about the representation of words in memory. In: V. A. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged* (pp. 101–114). New York: Academic Press.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics, 31,* 373–405.

Hirsh, I. J. (1988). Auditory perception and speech. In: R. C. Atkinson, R. J. Hermstein, G. Lindzey, & R. D. Luce (Eds), *Stevens' handbook of experimental psychology, volume I: Perception and motivation* (pp. 377–408). New York: Wiley.

Hollich, G., Newman, R. S., & Jusczyk, P. W. (2005). Infants' use of synchronized visual information to separate streams of speech. *Child Development, 76,* 598–613.

Hollien, H., & Klepper, B. (1984). The speaker identification problem. In: R. E. Rieber (Ed.), *Advances in forensic psychology and psychiatry* (Vol. 1, pp. 87–111). Norwood, NJ: Ablex.

Hollien, H., Majewski, W., & Doherty, E. T. (1982). Perceptual identification of voices under normal, stress and disguise speaking conditions. *Journal of Phonetics, 10,* 139–148.

Holst, E. von, & Mittelstaedt, H. (1950). Das reafferenzprinzip. Wechselwirkung zwischen zentralnervensystem and peripherie. [The reafference principle. Interaction between the central nervous system and the periphery.] *Naturwissenschaften, 37,* 464–476.

Honda, K. (1996). Organization of tongue articulation for vowels. *Journal of Phonetics, 24,* 39–52.

Houde, J. F., & Jordan M. I. (1998). Sensorimotor adaptation in speech perception. *Science, 279,* 1213–1216.

House, A. S., & Stevens, K. N. (1999). A longitudinal study of speech production, I: General findings. *Speech Communication Group Working Papers, XI,* 21–41.

Howell, P., & Darwin, C. J. (1977). Some properties of auditory memory for rapid formant transitions. *Memory & Cognition, 5,* 700–708.

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America, 97,* 553–562.

Jakobson, R., & Halle, M. (1956). *Fundamentals of language.* The Hague: Mouton & Co. Printers.

Johnson, K. (2005). Speaker normalization in speech perception. In: D. B. Pisoni, & R. E. Remez (Eds), *The handbook of speech perception* (pp. 363–389). Malden, MA: Blackwell.

Johnson, K., & Azara, M. (2000). The perception of personal identity in speech: Evidence from the perception of twins' speech. Unpublished manuscript.

Johnson, K., & Mullennix, J. W. (1997). *Talker variability in speech processing*. San Diego: Academic Press.

Johnson, M. K., Foley, M. A., & Leach, K. (1988). The consequences for memory of imagining in another person's voice. *Memory & Cognition, 16*, 337–342.

Jones, J. A., & Munhall, K. G. (2000). Perceptual calibration of *F*0 production: Evidence from feedback perturbation. *Journal of the Acoustical Society of America, 108*, 1246–1251.

Jones, J. A., & Munhall, K. G. (2003). Learning to produce speech with an altered vocal tract: The role of auditory feedback. *Journal of the Acoustical Society of America, 113*, 532–543.

Joos, M. (1948). Acoustic phonetics. *Language, 24* (supplement), 1–137.

Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Kent, R. D., & Chial, M. R. (2002). The scientific basis of expert testimony on talker identification. In: D. L. Faigman, D. H. Kaye, M. J. Saks, & J. Sanders (Eds), *Modern scientific evidence (Vol. 2*, pp. 590–630). Eagan, MN : West.

Kersta, L. G. (1962). Voiceprint identification. *Nature, 196*, 1253–1257.

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

Kohler, W. (1910). Akustische Untersuchungen, II. [Acoustic investigations]. *Zeitschrift fur Psychologie mit Zeitschrift fur Angewandte Psychologie und Charakterkunde, 58*, 59–140.

Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology, 38*, 618–625.

Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science, 1*, 113–114.

Kreiman, J. (1997). Listening to voices: Theory and practice in voice perception research. In: K. Johnson, & J. R. Mullenix (Eds), *Talker variability in speech* (pp. 85–108). New York: Academic Press.

Kreiman, J., Vanlancker-Sidtis, D., & Gerratt, B. R. (2005). Perception of voice quality. In: D. B. Pisoni, & R. E. Remez (Eds), *Handbook of speech perception* (pp. 228–362). Oxford: Blackwell.

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

Kühnert, B., & Nolan, F. (1999). The origin of coarticulation. In: W. J. Hardcastle, & N. Hewitt (Eds), *Coarticulation: Theory, data and techniques* (pp. 7–30). Cambridge: Cambridge University Press.

Labov, W. (1966). *The social stratification of english in New York City.* Washington, DC: Center for Applied Linguistics.

Labov, W. (1974). Linguistic change as a form of communication. In: A. Silverstein (Ed.), *Human communication: Theoretical explorations* (pp. 221–256). Hillsdale, NJ: Lawrence Erlbaum Associates.

Labov, W. (1986). Sources of inherent variation in the speech process. In: J. S. Perkell, & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 402–425). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Lachs, L., & Pisoni, D. B. (2004). Cross-modal source information and spoken word recognition. *Journal of Experimental Psychology: Human Perception & Performance, 30,* 378–396.

Lane, H. L., Catania, A. C., & Stevens, S. S. (1961). Voice level: Autophonic scale, perceived loudness and effects of sidetone. *Journal of the Acoustical Society of America, 33,* 160–167.

Lane, H., & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research, 14,* 677–709.

Lashley, K. S. (1951). The problem of serial order in behavior. In: L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior: The Hixon symposium* (pp. 112–136). New York: Wiley

Laver, J., (1980). *The phonetic description of voice quality.* Cambridge: Cambridge University Press.

Liberman, A. M. (1957). Some results of research on speech perception. *Journal of the Acoustical Society of America, 29,* 117–123.

Liberman, A. M. (1996). Introduction: Some assumptions about speech and how they changed. In: A. M. Liberman (Ed.), *Speech: A special code* (pp. 1–44). Cambridge, MA: MIT Press.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74,* 421–461.

Liberman, A. M., Delattre, P., & Cooper, F. S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *The American Journal of Psychology, 65,* 497–516.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1–36.

Liebenthal, E., Binder, J. R., Piorkowski, R. L., & Remez, R. E. (2003). Short-term reorganization of auditory analysis induced by phonetic experience. *Journal of Cognitive Neuroscience, 15,* 549–558.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech, 6,* 172–187.

Liénard, J. -S., & Di Benedetto, M. -G. (1999). Effect of vocal effort on spectral properties of vowels. *Journal of the Acoustical Society of America, 106,* 411–422.

Lindblom, B. (1996). Role of articulation in speech perception: Clues from production. *Journal of the Acoustical Society of America, 99*, 1683–1692.

Lombard, E. (1911). Le signe de l'elevation de la voix. [The sign of a rising voice.] *Annales des Maladies de L'oreille et du Larynx, 37*, 101–119.

Luce, P. A., Goldinger, S. D., Auer, E. T., Jr., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception & Psychophysics, 62*, 615–625.

MacNeilage, P. F. (1970). Motor control of serial ordering of speech. *Psychological Review, 77*, 182–196.

Magnuson, J. S., & Nusbaum, H. C. (2004). Acoustic differences, listener expectations, and the perception of talker differences. Unpublished manuscript.

Markel, J. D., & Davis, S. B. (1979). Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base. *IEEE Transactions on Acoustics, Speech & Signal Processing, ASSP-27*, 74–82.

Massaro, D., & Stork, D. (1998). Speech recognition and sensory integration: A 240 year old theorem helps explain how people and machines can integrate auditory and visual information to understand speech. *American Scientist, 86*, 236–244.

McFarland, D. H., Baum, S. R., & Chabot, C. (1996). Speech compensation to structural modifications of the oral cavity. *Journal of the Acoustical Society of America, 100*, 1093–11104.

McGehee, F. (1937). The reliability of the identification of the human voice. *Journal of General Psychology, 17*, 249–271.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.

McLennan, C. T., Luce, P. A., & Charles-Luce, J. (2003). Representation of lexical form. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29*, 539–553.

Miller, G. A. (1951). *Language and communication*. New York: McGraw-Hill.

Miller, G. A. (1965). Some preliminaries to psycholinguistics. *American Psychologist, 20*, 15–20.

Miller, J. L., & Volaitis, L. E. (1989). Effects of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics, 46*, 505–512.

Modell, J. D., & Rich, G. J. (1915). A preliminary study of vowel qualities. *American Journal of Psychology, 26*, 453–456.

Morsella, E., Romero-Canyas, R., Halim, J., & Krauss, R. M. (2002). Judging social identity from voice and photograph. Poster presented at meetings of Society for Personality and Social Psychology.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics, 58*, 351–362.

Natale, M. (1975). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality & Social Psychology, 32*, 790–804.

Navratil, J., Kleindienst, J,. & Maes, S. (2000). An instantiable speech biometrics module with natural language interface: Implementation in the telephony environment. *Proceedings of the IEEE ICASSP*-2000, Turkey, June, 2000.

Nolan, F. (1983). *The phonetic bases of speaker recognition.* Cambridge: Cambridge University Press.

Nolan, F., & Oh, T. (1996). Identical twins, different voices. *Forensic Linguistics, 3*, 39–49.

Nygaard, L., Sommers, M., & Pisoni, D. (1992). Effects of speaking rate and talker variability on the representation of spoken words in memory. In: J. Ohala (Ed.), *Proceedings of the international conference on spoken language processing,* Edmonton, Alberta, University of Alberta Press (pp. 591–594).

Nygaard, L., Sommers, M., & Pisoni, D. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5*, 42–46.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America, 119*, 2382–2393.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*, 175–184.

Picheny, M. (2004). Achieving superhuman speech recognition: Via heredity or environment? *Minutes of the Columbia University Seminar on Language & Cognition*, January 29, 2004.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1985). Speaking clearly for the hard of hearing, In: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research, 28*, 96–103.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral & Brain Sciences, 27*, 169–226.

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In: K. Johnson, & J. W. Mullenix (Eds), *Talker variability in speech processing* (pp. 9–32). San Diego: Academic Press.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics, 15*, 285–290.

Remez, R. E. (2001). The interplay of phonology and perception considered from the perspective of organization. In: E. V. Hume, & K. A. Johnson (Eds), *The role of speech perception phenomena in phonology* (pp. 27–52). New York: Academic Press.

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1996). Phonetic sensitivity and individual recognition: Notes on system architecture. Paper presented at the symposium on speaker characteristics: Effects of gestural variability on speech perception. Third Joint Meeting, Acoustical Societies of America and Japan, Honolulu, HI, December 2.

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 651–666.

Remez, R. E., Pardo, J. S., Piorkowski, R. L., & Rubin, P. E. (2001). On the bistability of sinewave analogues of speech. *Psychological Science*, *12*, 24–29.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review*, *101*, 129–156.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, *212*, 947–950.

Richardson-Klavehn, A., & Bjork, R. A. (1988). Measures of memory. In: M. R. Rosenzweig, & L. W. Porter (Eds), *Annual review of psychology*, (Vol. 39, pp. 475–543). Palo Alto: Annual Reviews.

Rizzolati, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, *27*, 169–192.

Rosen, S. M., Fourcin, A. J., & Moore, B. C. J. (1981). Voice pitch as an aid to lipreading. *Nature*, *291*, 150–152.

Rosenblum, L. D. (2005). Primacy of multimodal speech perception. In: D. B. Pisoni, & R. E. Remez (Eds), *The handbook of speech perception* (pp. 51–78). Oxford: Blackwell.

Rosenblum, L. D., & Gordon, M. S. (2001). The generality of specificity: Some lessons from audiovisual speech. *Behavioral and Brain Sciences*, *24*, 239–240.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics*, *31*, 307–314.

Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, *25*, 421–436.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, *99*, 143–165.

Shannon, R. V., Zeng, F. -G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*, 303–304.

Shepard, C. A., Giles, H., & Le Poire, B. A. (2001). Communication accommodation theory. In: W. P. Robinson, & H. Giles (Eds), *The new handbook of language & social psychology* (pp. 33–56). New York: Wiley.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*, 87–90.

Stevens, K. N. (1990). Lexical access from features. A paper presented at the workshop on speech technology for man–machine interaction, Bombay, India.

Stevens, K. N., & Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In: P. D. Eimas, & J. L. Miller (Eds), *Perspectives on the study of speech* (pp. 1–38). Hillsdale: Lawrence Erlbaum Associates.

Studdert-Kennedy, M., & Goldstein, L. (2003). Launching language: The gestural origin of discrete infinity. In: M. H. Christiansen, & S. Kirby (Eds), *Language evolution: The states of the art* (pp. 235–254). Oxford: Oxford University Press.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212–215.

Svirsky, M. A., Robbins, A. M., Kirk, K. I., Pisoni, D. B., & Miyamoto, R. T. (2000). Language development in profoundly deaf children with cochlear implants. *Psychological Science, 11*, 153–158.

Tartter, V. C. (1980). Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception & Psychophysics, 27*, 24–27.

Uchanski, R. M. (2005). Clear speech. In: D. B. Pisoni, & R. E. Remez (Eds), *The handbook of speech perception* (pp. 207–235). Oxford: Blackwell.

Vallabha, G. K., & Tuller, B. (2004). Perceptuomotor bias in the imitation of steady-state vowels. *Journal of the Acoustical Society of America, 116*, 1184–1197.

Vick, J. C., Lane, H., Perkell, J. S., Matthies, M. L., Gould, J., & Zandipour, M. (2001). Covariation of cochlear implant users' perception and production of vowel contrasts and their identification by listeners with normal hearing. *Journal of Speech, Language and Hearing Research, 44*, 1257–1267.

Vitevitch, M. S. (2003). Change deafness: The inability to detect changes between two voices. *Journal of Experimental Psychology: Human Perception & Performance, 29*, 333–342.

Watson, C. S., Qiu, W. W., Chamberlain, M. M. , & Li, S. (1996). Auditory and visual speech perception: Confirmation of a modality-independent source of individual differences in speech recognition. *Journal of the Acoustical Society of America, 100*, 1153–1162.

Weijer, J. Van de. (1997). Language input to a prelingual infant. In: A. Sorace, C. Heycock, & R. Shillcock (Eds). *Proceedings of the GALA '97 Conference on Language Acquisition* (pp. 290–293). Edinburgh, Scotland: GALA.

Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory & Language, 31*, 183–194.

Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *Journal of the Acoustical Society of America, 52*, 1238–1250.