Journal of Experimental Psychology:
Human Perception and Performance
2007, Vol. 33, No. 2, 391–409

0096-1523/07/$12.00 DOI: 10.1037/0096-1523.33.2.391

Acoustic Differences, Listener Expectations, and the Perceptual Accommodation of Talker Variability

James S. Magnuson
University of Connecticut and Haskins Laboratories

Howard C. Nusbaum
University of Chicago

Two talkers' productions of the same phoneme may be quite different acoustically, whereas their productions of different speech sounds may be virtually identical. Despite this lack of invariance in the relationship between the speech signal and linguistic categories, listeners experience phonetic constancy across a wide range of talkers, speaking styles, linguistic contexts, and acoustic environments. The authors present evidence that perceptual sensitivity to talker variability involves an active cognitive mechanism: Listeners expecting to hear 2 different talkers differing only slightly in average pitch showed performance costs typical of adjusting to talker variability, whereas listeners hearing the same materials but expecting a single talker or given no special instructions did not show these performance costs. The authors discuss the implications for understanding phonetic constancy despite variability between talkers (and other sources of variability) and for theories of speech perception. The results provide further evidence for active, controlled processing in real-time speech perception and are consistent with a model of talker normalization that involves contextual tuning.

Keywords: speech perception, phonetic constancy, talker normalization

A fundamental theoretical problem that remains unsolved despite decades of speech research is the lack of invariance between the speech signal and perceptual categories. Depending on phonetic context, talker characteristics, or speaking rate, one acoustic pattern may correspond to several different phonemes, and one phonemic percept may result from several different acoustic patterns (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Peterson & Barney, 1952). Despite this lack of invariance, listeners routinely experience *phonetic constancy*: They successfully recognize the variable speech input as the intended phonetic categories, just as they have a stable percept of color across variation in ambient illumination or the size of a person across retinal image differences.

Over the years, much of the research on the lack-of-invariance problem has focused on the role of acoustic–phonetic variability as a consequence of coarticulation among phonetic segments (e.g., Liberman, 1986; Liberman et al., 1967). For example, a particular consonant release burst is heard as /p/ or /k/ depending on whether the following context is /i/ or /a/ (Liberman, Delattre, & Cooper, 1952). In addition, the formant transition pattern corresponding to any particular consonant varies depending on context (Delattre, Liberman, & Cooper, 1955).

James S. Magnuson, Department of Psychology, University of Connecticut, and Haskins Laboratories, New Haven, Connecticut; Howard C. Nusbaum, Department of Psychology, University of Chicago.

This research was supported by National Institute of Deafness and Other Communication Disorders Grant DC-005765 to James S. Magnuson and National Institute of Child Health and Human Development Grant HD-01994 to Haskins Laboratories. We thank Inge-Marie Eigsti, Ted Strauss, and Harlan Harris for helpful discussions.

Correspondence concerning this article should be addressed to James S. Magnuson, Department of Psychology, 406 Babbidge Road, Unit 1020, University of Connecticut, Storrs, CT 06269-1020. E-mail: james.magnuson@uconn.edu

When differences in talker and speaking rate are considered, the problem of lack of invariance becomes considerably worse, because these factors provide additional sources of many-to-many mappings. Across talkers, the same acoustic pattern can correspond to different perceptual categories, whereas different talkers' productions of the same phoneme can differ acoustically (Dorman, Studdert-Kennedy, & Raphael, 1977; Peterson & Barney, 1952). Similarly, one acoustic pattern may be heard as /b/ at one speaking rate but as /w/ at a slower rate (J. L. Miller & Baer, 1983). Thus, different sources of variability contribute to the nondeterministic mapping of acoustic patterns onto phonetic categories (cf. Nusbaum & Magnuson, 1997).

In spite of the similar consequences for theories of speech perception, different sources of acoustic–phonetic variability in speech have typically been investigated independently. Phonetic context variability has generally been treated as the primary problem (e.g., Liberman et al., 1967), and factors like talker differences and speaking rate variability have been treated as secondary complications. On the one hand, this is a useful simplifying assumption; examination of one type of variability is more tractable when the others are held constant. On the other hand, it seems to suggest that different types of variability pose distinct perceptual challenges. Indeed, this appears to follow from observations that each type of acoustic–phonetic variability is marked by different kinds of signal properties grounded in the physics of articulation (e.g., burst locus and formant transition patterns for phonetic context variability [Delattre et al., 1955]; formant frequency relationships for talker differences [Syrdal & Gopal, 1986]; acoustic segment durations for speaking rate [J. L. Miller & Liberman, 1979]).

Studies of phonetic context variability typically focus on relatively local influences, such as the effects of coarticulation among adjacent segments (e.g., Lotto, Kluender, & Holt, 1997; Mann, 1980). These studies tend to focus on the way in which temporally proximal sets of cues influence phonetic perception. In contrast,

research on talker and speaking rate variability has often examined the effects of much larger temporal contexts, such as an antecedent carrier phrase (e.g., Ladefoged & Broadbent, 1957, for talker differences; Gordon, 1988, for speaking rate variability), although the durations of adjacent segments have also been examined (e.g., J. L. Miller & Liberman, 1979).

In spite of these differences in the research approach, similar theoretic assumptions arise in studying these different forms of variability, unified by the goal of finding a deterministic mapping from acoustic patterns to perceptual categories. One underlying assumption is that there are phonetic invariants but that typical methods of speech analysis do not reveal the actual invariant information, whether in acoustic cues (e.g., Kewley-Port, 1983; Stevens & Blumstein, 1978, 1981) or in causal event structures directly informing the listener of the movements of articulators (i.e., direct realism: Best, 1994; Fowler, 1989). A distinct but conceptually related assumption is that a sufficient sampling of the acoustic distribution of utterances might provide a statistical basis for phonetic stability (e.g., Diehl, Lotto, & Holt, 2004; Goldinger, 1998). These assumptions are predicated on the notion that there is some kind of invariant basis to phonetic perception but that the true nature of the invariance is obscured by the way in which the acoustic patterns of speech are analyzed. In contrast, motor theory (Liberman et al., 1967; Liberman & Mattingly, 1985) accepts that there are no true or underlying acoustic cues to phonetic categories. Rather, motor theory proposes that listeners use knowledge of speech production (embedded in a speech-specific mechanism) to perceptually recover the articulatory gestures used in speaking. Thus, invariance lies in the motor behavior rather than the acoustic information.

However, these assumptions all share the idea that the process of mapping from acoustic cues to phonetic categories is direct and unmediated by cognitive processes. The classification of an external signal (acoustic cue or motor movement) is directly mapped or transformed passively to a corresponding phonetic category. Thus, the relationship from external signal to perceptual category is a deterministic process that essentially strips away irrelevant variability.

Even after decades of speech research, there are few direct tests of this view. Empirical support comes in the form of evidence for or against specific predictions, relatively limited in scope, that are associated with aspects of this view but do not test it. For example, evidence that nonhuman animals are sensitive to the covariation of acoustic information in speech (e.g., Lotto et al., 1997) is used to argue that articulatory knowledge is not necessary to explain phonetic context effects. However, this evidence cannot rule out the possibility that such knowledge is actually used by human listeners. Similarly, the effect of talker-specific details of spoken words on memory and responses (Goldinger, 1998) is used to argue that talker-specific information is not stripped away from the perception of those words. But this does not prove that holistic traces are the sole medium of speech perception (see the General Discussion). Many such studies focus on isolating one aspect of variability and demonstrating that the problem posed by this variability can be resolved, in principle, by some kind of direct mapping process that could be described as a bottom-up process or a "passive" mapping (see Nusbaum & Schwab, 1986).

We have argued previously (Nusbaum & Magnuson, 1997) that by its very definition, the lack of invariance in the mapping of acoustic patterns onto phonetic categories is computationally nondeterministic and, thus, cannot be resolved by any direct, deterministic mapping theory. The theoretical approaches that have been taken in the past have been based entirely on the pattern properties of the speech stimulus and have attempted to solve the lack-of-invariance problem by positing the use of specific knowledge (e.g., special invariant features, statistical distributions, articulatory gestures) that could provide a deterministic mapping of speech onto phonetic categories. However, analytic consideration of the nature of these direct mapping theories and the nondeterministic nature of the computational problem suggests that this is an untenable approach.

Thus, we argued that the nondeterministic mapping must be solved by mechanisms incorporating active control structures (Nusbaum & Magnuson, 1997). Unpacking this claim requires a brief discussion of the three fundamental components of computational mechanisms: representations, transformations, and control structures. If a deterministic mapping exists between input and output representations, or if the representations can be transformed to render a deterministic mapping, passive control structures (also called *open-loop* structures) suffice: The same input conditions result in the same series of transformations and outputs. For example, an open-loop movement like a ballistic saccade is planned in advance, and the movement is executed without on-the-fly adjustment. Such structures afford fast, automatic responses, but they are inflexible and context-invariant. With active control structures (also called *closed-loop* structures), the same input need not result in the same output. Instead, the system monitors and modifies its output in a context-sensitive way. A classic closed-loop example is a visually guided reach, in which visual feedback can be used to continuously adjust the movement as it is made. Thus, active control structures are generally more complex, but they afford adaptive, flexible, context-dependent responses. On the active control view of speech processing, the variability in speech does not obscure the underlying phonetic code but lawfully reflects properties of the message source—the talker and other context. Phonetic constancy is achieved by adjusting interpretation of the signal as a function of bottom-up and top-down constraints. We call this the *active control hypothesis.*

In this article, we focus on the issue of talker variability and explore how much acoustic difference between talkers is sufficient for the perceiver to show effects of talker variability (e.g., Nusbaum & Morin, 1992). We begin by briefly reviewing the primary findings associated with accommodating talker variability. We show that there is growing evidence for active control but also that some basic facts about talker variability have yet to be examined.

First, the basic challenge of talker variability is that the relationship between acoustic patterns and phonetic categories depends on specific vocal characteristics of a talker. That is, there is a talker-dependent mapping between acoustic information and linguistic categories. For example, Peterson and Barney (1952) plotted the vowels of many adult male and female talkers and of child talkers in F1 × F2 space (i.e., the first two formants) and found substantial overlap between categories, such that one talker's production of a vowel like /ɪ/ might have the same formants as another talker's production of /ɛ/, whereas the two talkers' productions of /i/ could be quite different. This is a classic example of the many-to-many mapping between acoustics and phonetics: Two vowel categories may be produced by different talkers with very similar acoustic patterns, whereas very different acoustic patterns produced by different talkers may be intended

and perceived as the same vowel. However, listeners have little trouble interpreting acoustic patterns in talker-appropriate ways, even from isolated words (Peterson & Barney, 1952). Although few studies approach the scope and breadth of this particular study, Johnson, Ladefoged, and Lindau (1993) reported that talkers vary rather dramatically in the articulatory gestures they use to produce acoustically similar segments. Talker-dependent mappings between acoustic patterns and consonant categories have also been reported. For example, even when dialect, sex, and emotional state are relatively controlled, two talkers can differ in their production of acoustic cues to place of articulation such that one talker can rely more on release burst spectral shape, whereas another will rely more on transitions (e.g., Dorman et al., 1977). Thus, listeners cannot weight available cues equally for all talkers, because the information afforded by different cues may provide different phonetic information depending on the talker. The listener must figure out something about the way these talkers produce speech so as to recognize the intended phonetic information and use coarticulatory cues. However, to do so, the listener must perceive that such information is needed. How much of a difference in talker vocal characteristics is needed for the listener to encode this acoustic information differently?

The simplest theories of how listeners achieve phonetic constancy despite talker variability are *intrinsic normalization* theories (cf. Ainsworth, 1975; Nearey, 1989) or structural estimation theories (Nusbaum & Morin, 1992). On this view, any sample of speech is "self-normalizing"; that is, any utterance contains sufficient information to properly categorize it via a normalizing transformation. For example, Syrdal and Gopal (1986) proposed that F0 (fundamental frequency) and F3 (the third formant) could be used to rescale F1 and F2, the primary cues to vowel identity, although such a rescaling does not eliminate the effects of talker variability entirely (similarly, the duration of a vowel can provide the basis for estimating speaking rate to calibrate perception of a stop or semivowel consonant; J. L. Miller & Liberman, 1979). Shankweiler, Strange, and Verbrugge (1977) proposed that the dynamic cues in consonant formant transitions (coarticulated with an adjacent vowel) may provide talker-independent information about vowel identity. This approach was supported by tests of computational models and related later work (e.g., Strange, 1989), suggesting that individual samples of speech indeed contain considerable talker-independent phonetic information. However, intrinsic normalization cannot account for the next two phenomena.

The second major talker variability finding is that there is a processing cost when there is talker variability in a set of utterances. Changing from one talker to another in experimental settings can impair accuracy and/or response time. Speech perception in noise is less accurate and slower when there is talker variability in a set of utterances compared with when listeners hear utterances produced by a single talker (Creelman, 1957; Magnuson, Nusbaum, & Akahane-Yamada, 2007; Mullennix, Pisoni, & Martin, 1989). Recognition of vowel, syllable, and spoken word targets is reliably slower when there is talker variability compared with recognition of the speech of a single talker, and this slowing of responses interacts with working-memory load (Nusbaum & Morin, 1992; see also Mullennix & Howe, 1999). Talker consistency promotes better performance than does talker variability in shadowing speech (Goldinger, 1998). Even serial recall of lists of words spoken by different talkers is worse than recall of lists of words spoken by single talkers (Martin, Mullennix, Pisoni, & Summers, 1989). Intrinsic normalization does not predict a cost associated with talker changes, because the normalizing transformation is carried out for every sample of speech. This, along with the next finding, motivates an alternative theory of normalization.

The third major finding is that the context of prior speech influences recognition of a target utterance. For example, following Joos's (1948) suggestion that antecedent information about a talker's vowel space might provide a context for subsequent vowel perception, Ladefoged and Broadbent (1957) manipulated the vocal characteristics of a synthesized carrier phrase (by shifting formant frequencies) to simulate different talkers' vowel spaces, and they observed significant changes in perception of a constant set of vowel targets that depended on the characteristics of the preceding carrier phrase (see Gordon, 1988, for analogous effects based on the speaking rate of an antecedent context for a voicing decision). Thus, information about a talker's vocal tract perceived prior to a particular vowel sample is used by listeners in mapping the acoustics of the sample onto phonemic categories.

Effects of preceding context have motivated *extrinsic normalization* theories of talker variability (cf. Ainsworth, 1975; Nearey, 1989). On this view, listeners use intrinsic cues but also improve the talker-specific mapping as more speech is heard. For example, Gerstman (1968) tested a computational model that achieved good performance across talkers when given information about the formant parameters for a talker's point vowels that were used to rescale subsequent speech (although it remains to be seen whether this simple approach can be extended to fluent speech). On Nusbaum and Morin's (1992) contextual tuning theory, a change in talker triggers normalization procedures that operate until a stable mapping between the talker and internal phonetic categories is achieved. The stable mapping is then maintained until a talker change is indicated acoustically (e.g., by large changes in F0) or more implicitly (e.g., via failures of lexical access). Such accounts naturally explain effects of prior context. Extrinsic theories also provide an account for processing costs following talker changes, because they exploit stability in talker characteristics and only perform normalization after talker changes.

Nusbaum and Morin (1992) also reported results that are consistent with the active control hypothesis. First, perceived changes in the talker affect the way listeners attend to subsequent speech. Nusbaum and Morin reported that when the speech was produced by a single talker, with formants above F2 filtered, whispering the vowels or combining whispering and filtering had no effect on recognition accuracy. However, when there was talker variability, elimination of F0 and F3 and higher formants impaired accuracy. This suggests that listeners only relied on these cues for phonetic recognition after recognizing a change in talker (implicitly or explicitly). Second, talker variability interacts with working memory load to affect speech perception (Nusbaum & Morin, 1992). When listeners are given a list of visually presented numbers to hold in working memory while recognizing spoken syllables, there is no effect on speech recognition performance of varying the length of the list (amount of memory load) when the speech is produced by one talker. However, when there is talker variability, increasing working-memory load significantly slows speech recognition, suggesting that talker variability itself may place demands on working memory (cf. Navon, 1984). This interaction is also typical of controlled processing (Schneider & Shiffrin, 1977),

as Nusbaum and Magnuson (1997) discussed in detail with respect to this finding (see also Nusbaum & Schwab, 1986).

Recently, Wong, Nusbaum, and Small (2004) reported neural evidence consistent with the hypothesis that talker changes engage controlled attentional processing. They compared word-monitoring performance in blocked- and mixed-talker conditions and measured changes in neural activity using functional MRI (fMRI). The mixed-talker condition resulted in increased activity in cortical areas associated with speech processing under resource-demanding conditions (posterior superior temporal gyrus; e.g., Just, Carpenter, Keller, Eddy, & Thulborn, 1996) and shifts of attention to spectral information (superior parietal cortex; e.g., Liu, Slotnick, Serences, & Yantis, 2003; Posner, 2003).

Although our active control hypothesis invokes a computationally distinct alternative to conventional approaches to lack of invariance, it has developed naturally from normalization approaches to the problem. Another alternative to conventional theories are *nonanalytic episodic* theories (Goldinger, 1998; Pisoni, 1997). This view is based on a radically different approach to the representations underlying speech perception and lexical access intended to approximate a deterministic mapping. While we defer discussion of the details of these and related theories and the results that motivate them until the General Discussion, the key idea is that by preserving holistic (hence, nonanalytic) exemplar traces of speech events, statistical clustering of the traces would provide a basis for phonetic constancy without the need for an explicit mechanism to accommodate talker variability or other forms of context dependence. The present experiments provide opportunities to evaluate the claim that perceptual accommodation of talker variability is not needed.

As this brief overview shows, talker variability tends to decrease the efficiency of speech perception. However, there is a very basic question that has yet to be addressed by any of the research on talker variability: What constitutes a difference between talkers sufficient to produce these variability effects? Experiments 1–3 were designed to investigate how listeners detect a change in talker characteristics. Experiment 4 returned to a deeper theoretical issue. The effects on attention and working memory found by Nusbaum and Morin (1992) suggest that active control structures may underlie the perceptual accommodation of talker variability. In Experiment 4, we tested whether unequivocally top-down information modulates responses to talker variability. The results point to the nature of the control structures underlying speech perception as the most promising avenue for solving the lack-of-invariance problem.

Experiment 1

How do acoustic differences between talkers' vocal characteristics relate to the effects of talker variability on speech perception? As noted above, talker variability effects are manifested in several ways, such as reduced recognition accuracy and slower recognition. Talker normalization theories assume that when there is a talker change, listeners must figure out how to map the new talker's acoustic patterns onto their intended phonetic interpretations, and determination of the new mapping will slow recognition. In comparison, episodic-trace models such as Goldinger's (1998) deny normalization and, instead, attribute talker variability effects to the simultaneous encoding of indexical and phonetic information.

From either perspective, as long as the talker is constant, performance is maintained; the change in talker generates the effect on recognition performance. However, neither perspective makes strong claims about how much change between talkers is necessary for talker variability effects to be obtained. For the episodic models, in principle, if there is a discriminable change in a talker's voice, listeners will have to attend to both phonetic and indexical dimensions, thus slowing phonetic recognition. This means that whether the indexical properties change or, for example, the vowel space changes, there should be a variability effect. However, normalization models could make a different prediction. Because normalization is only necessary to constrain phonetic interpretation, normalization theories might predict that talker variability effects will be observed only if there is a difference between talkers that is sufficient to produce a many-to-many mapping. That is, if two talkers' vowel spaces are sufficiently similar, talker variability effects might not be observed, even if the talkers' voices differ in other qualities.

Normalization theories would predict that talker variability effects result from the increased ambiguity in acoustic-to-phonetic mapping and the performance costs of adjusting to the acoustic–phonetic space of the new talker (Nusbaum & Magnuson, 1997). However, a talker's vocal characteristics are correlated with other aspects of that talker's speech. Thus, it would not be surprising if normalization operated by a satisficing (rather than optimal) approach of engaging in normalization whenever any detectable acoustic change between talkers is perceived. Although this more conservative strategy might lead to a greater average slowing of recognition performance (through unnecessary normalization operations), it would promote recognition accuracy.

Do all pairs of discriminably different talkers produce talker variability effects when their speech is presented to listeners? Does the size of the talker variability effect depend on the acoustic difference between different talkers? Experiment 1 represents an extension of an experiment reported by Nusbaum and Morin (1992), using the same materials but additional subjects. We used speeded spoken target monitoring to examine whether the acoustic difference between talkers in a mixed-talker condition predicts the magnitude of normalization effects. A talker variability effect is operationalized as reliable slowing of speech recognition (or reliable decreases in accuracy) under conditions of talker change compared with conditions of talker stability (i.e., mixed- vs. blocked-talker conditions). In this experiment, we compared mixed-talker conditions with pairs of talkers of the same sex or different sexes. Within each pair of talkers, the voices were certainly discriminable, although different pairs were not equally discriminable. If talker variability effects depend solely on discriminability of voices, all pairs should have produced similar effects relative to single-talker conditions. If talker variability effects depend on whether the difference between talkers is sufficient to produce a need for normalization tuning, then different pairs may produce different results.

Method

Subjects. Two groups of 23 subjects were recruited from the University of Chicago community. All subjects were native speakers of American English and reported that they had no history of

hearing or speech disorders. We paid subjects for their participation and debriefed them when they completed the experiment.

Stimuli. The stimuli consisted of the vowels /i/, /I/, /u/, and /U/ (used as targets and distractors) and /e/, /æ/, /a/, and /Λ/ (used only as distractors). One token of each vowel was produced by each of two male and two female talkers. We measured the formant frequencies of the steady-state portions of each vowel for all four talkers. Although the experimenters found the male and female pairs easily discriminable (they were familiar with all four talkers), a surprising observation followed from the vowel space measures. The male talkers differed substantially, but the female talkers had nearly identical vowel spaces as well as average pitches of approximately 200 hz. Formant centers measured by hand from spectrograms of the steady-state portions of the target vowels for each talker, along with /æ/ and /a/ to more fully describe the vowel spaces, are shown in Table 1. The items were padded with silence to make all items 500 ms in duration, so as to allow for a constant stimulus onset asynchrony (SOA).

Procedure. A speeded target-monitoring task was used, and hit rate, false alarm rate, and response times were calculated. Subjects were presented with an orthographic form of a target vowel on a computer display and instructed to press the space bar whenever they heard the target vowel they saw on the screen. On each trial, subjects heard a sequence of 16 vowels separated by 250 ms of silence (making the total SOA 750 ms). Four targets were randomly placed at ordinal positions between the 1st and 16th stimuli, and they were separated by at least 1 distractor. On each trial, subjects listened for 1 target from the set /i/, /I/, /u/, and /U/. The twelve distractors were randomly selected from the full set of stimuli, excluding the trial's target.

The two groups of subjects each heard a blocked-talker condition and two mixed-talker conditions, with block order counterbalanced. Each block consisted of 12 trials (each trial being a 16-vowel sequence with 4 of the stimuli being targets). There were two mixed-talker blocks and two blocked-talker blocks, with order counterbalanced between subjects. For one group, the mixed-talker conditions contained two talkers of the same sex, with two female talkers in one block and two male talkers in the other (with the two talkers mixed within each block). The other group heard pairs of different-sex talkers in the mixed blocks. In one mixed-talker block, one male and one female talker were combined, and the other two talkers were combined in the second mixed-talker block. For every talker a subject heard in the mixed-talker condition, there was a corresponding blocked-talker set of trials (on which all the items were produced by that talker). For each subject, each target produced by each talker occurred four times in the blocked-talker condition and four times in the mixed-talker condition. Thus, for each talker, there were 16 mixed-talker condition trials and 16 blocked-talker condition trials. In the blocked-talker condition, trials (as well as items within trials) were blocked by talker, and talker order was counterbalanced.

Table 1
Vowel Spaces (in Hz) of the Natural Talkers Used in Experiment 1

		Vowel					
Talker	Formant	/i/	/I/	/a/	/æ/	/U/	/u/
Female 1	F1	280	415	970	940	485	355
	F2	2,750	2,170	1,500	2,040	1,255	1,080
Female 2	F1	350	420	900	850	480	385
	F2	2,650	2,280	1,310	2,010	1,300	1,350
Male 1	F1	380	425	765	585	510	365
	F2	2,105	1,875	1,250	1,800	1,180	960
Male 2	F1	210	375	730	550	385	295
	F2	2,185	1,890	1,175	1,900	1,100	775

Results and Discussion

We measured response time and accuracy for responses to targets (hits) and distractors (false alarms). Response times were measured from stimulus onset, and responses that occurred less than 150 ms after the onset of a stimulus were scored as responses to the preceding item. We conducted analyses of variance (ANOVAs) on hit rate, false alarm rate, and response time for two factors with two levels each: talker sex homogeneity (same or different) and talker condition (blocked or mixed). Recall that for each talker a subject heard in the mixed-talker condition, there was a corresponding blocked-talker set of trials on which all stimuli were produced by that single talker. Thus, for example, for the two male talkers, the blocked-talker condition consisted of separate blocks of blocked-talker trials with each talker, whereas the mixed-talker condition consisted of the same number of trials per talker on which the talker switched randomly from one male talker to the other.

Hit rates were slightly higher when talker sexes were different (.94, SD = .07) than when they were the same (.92, SD = .08), but the main effect of talker sex was not significant, $F(1, 44) = 1.01$, $p = .319$, $\omega^2 = 0$. The main effect of talker condition (blocked by talker vs. mixed talkers) was significant, $F(1, 44) = 4.27$, $p = .045$, with higher hit rates in the blocked-talker condition (.94, SD = .07) than in the mixed-talker condition (.91, SD = .08), although the effect size was negligible ($\omega^2 = .002$). The Talker Sex Homogeneity × Talker Condition interaction nearly reached significance for hit rates, $F(1, 44) = 3.40$, $p = .072$, but the effect size was also negligible ($\omega^2 = .001$), and we did not explore this interaction further.

A similar pattern held for false alarm rates. The main effect of sex homogeneity was not significant, $F(1, 44) \leq 1$, with the mean false alarm rate being .02 in both conditions. There was a weak main effect of talker condition, $F(1, 44) = 11.76$, $p = .001$, $\omega^2 = .006$, with slightly more false alarms in the mixed-talker condition (.03, SD = .02) than in the blocked talker condition (.02, SD = .02). The Talker Sex Homogeneity × Talker Condition interaction was not significant, $F(1, 44) < 1$.

The pattern of response time results was more complex. The main effect of talker sex homogeneity was not significant, $F(1, 44) < 1$ (mean for same-sex pairs = 444 ms, SD = 43; mean for different sex pairs = 448 ms, SD = 48). There was a modest main effect of talker condition, $F(1, 44) = 25.26$, $p < .001$, $\omega^2 = .01$, with faster responses in the blocked-talker condition (437 ms, SD = 45) than in the mixed-talker condition (466 ms, SD = 42). There was also a weak Talker Sex Homogeneity × Talker Condition interaction, $F(1, 44) = 5.72$, $p = .021$, $\omega^2 = .002$, so we examined the simple effect of talker condition at each level of talker sex homogeneity. The effect was reliable for different-sex pairs, $F(1, 22) = 37.16$, $p < .001$, $\omega^2 = .07$ (mean blocked = 431

ms, SD = 44; mean mixed = 458 ms, SD = 40), but not for same-sex pairs, $F(1, 22) = 2.75$, $p = .111$, $\omega^2 = .003$ (mean blocked = 443 ms, SD = 47; mean mixed = 453 ms, SD = 45).

Given the vowel space differences observed in describing the stimuli—the two female talkers were more similar than were the two male talkers—we probed further by examining the simple effect of talker condition for each talker pair. The means in each talker condition are shown for each talker pair in Figure 1. Response times were reliably longer in the mixed-talker condition than in the blocked-talker condition for both different-sex pairs and for the two male talkers, $F(1, 22) > 14.00$, $p < .001$, $\omega^2 > .03$ (for all three pairs), but not for the two female talkers, $F(1, 22) < 1$. Note that similar effect sizes were found for the two different-sex pairs ($\omega^2 = .03$ or .04) and the pair of male talkers ($\omega^2 = .03$). This suggests that there is not a graded effect of talker differences; the substantial pitch and vowel space differences between the male and female talkers did not lead to more slowing than the more modest differences between the two male talkers.

Given that the experimenters subjectively found the female talkers as discriminable as the male talkers (they were familiar with all four talkers), the absence of a talker condition effect for the two female talkers suggests at least three implications for talker variability. First, the attentional demands that slow processing in the mixed condition may be a result of detection (whether explicit or not) of variability that may require normalization, such as a significant change in vowel space. That is, in both the blocked and mixed conditions, subjects begin contextual tuning on the basis of the characteristics of the talker who produced the first stimulus in the block. If the next stimulus is sufficiently similar, contextual tuning continues to build on the results of the first computation. Otherwise, contextual tuning must be restarted, thereby slowing recognition and increasing cognitive load. In the case of the female talkers, their vowel spaces may have been sufficiently similar for the perceptual framework computed for one to work for the other.

Second, blocked- vs. mixed-talker effects may indicate mandatory processing of talker characteristics (Goldinger, 1998; Nygaard & Pisoni, 1998). However, the lack of an effect for the two female talkers who seemed quite discriminable could be interpreted as not supporting this view. Of course, it is possible that the reason we did not find an effect with the female talkers was simply that the acoustic differences between the talkers were too small; listeners unfamiliar with the female voices may have failed to perceive the difference—especially given the short, isolated vowel stimuli used.

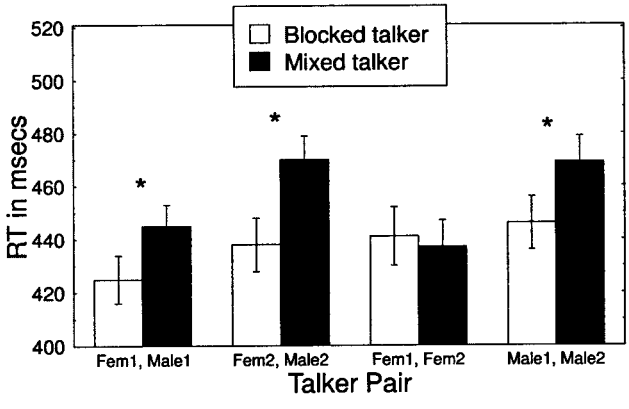

Figure 1. Response times (RTs) for each talker pair in blocked- and mixed-talker conditions in Experiment 1. Asterisks indicate statistically significant differences. Error bars represent standard error. Fem = female.

The third implication, which follows from the second, is that there is a nondeterministic mapping between acoustic cues and talkers, just as there is between acoustics and phonetic categories. That is, if the talkers were not distinct to listeners unfamiliar with their voices, despite being distinct to the experimenters, this would suggest that there is not a simple mapping from acoustic differences to talker differences; rather, listeners may need to learn what characteristics distinguish similar talkers.

However, we do not know much about what characteristics lead to significant talker differences (e.g., differences perceived as a change in talker identity and/or a change large enough that a new speech–phonetic category mapping is triggered). The lack of a variability effect with the two female talkers suggests that a small difference in vowel space is not sufficient to yield a talker condition effect. In Experiments 2 and 3, we began to isolate the types and magnitude of acoustic differences that lead to reliable performance differences associated with talker differences. In Experiment 4, we returned to the question of whether the mapping between acoustics and talkers is nondeterministic and to the nature of the control structures needed to account for perceptual constancy despite talker differences.

Experiment 2

Our initial interpretation of Experiment 1 is that a perceptible change in talker identity need not trigger normalization if the two talkers' vowel spaces are sufficiently similar. Experiment 2 used synthetic talkers to examine what sorts of acoustic variability triggers normalization. We manipulated two acoustic factors (F0—the basis for pitch—and vowel space) that have been proposed to be important for normalization (e.g., Syrdal & Gopal, 1986) and one other factor (amplitude) that can be linguistically relevant (e.g., prosodically) but that would have little correlation with talker differences. Several previous studies have found that amplitude does not pattern with linguistically relevant variation such as changes in talker or speaking rate. For example, word identification performance decreases when speaking rate is varied within a list but not when amplitude is varied (Sommers, Nygaard, & Pisoni, 1994), and preserving amplitude in new–old recognition memory tasks does not provide a benefit the way preserving rate or talker characteristics does (Bradlow, Nygaard, & Pisoni, 1999). However, the impact of amplitude variation in an online task like monitoring has not been tested. On the one hand, there is little reason to expect an effect of amplitude given previous null results in paradigms that often show parallel effects with online monitoring—although this may also mean an amplitude manipulation can provide a baseline for acoustic parameters that have little linguistic relevance. On the other hand, if slowing in the mixed-talker condition results from simple distraction, changes in any of our parameters in the mixed condition—including amplitude—should result in significant slowing. If instead the mechanisms that slow processing in mixed-talker conditions are triggered by acoustic changes indicative of talker changes, we should find slowing only for vowel space and F0 differences (because large changes in F0

are typically associated with changes in vowel space as well; Fujisaki & Kawashima, 1968; R. L. Miller, 1953; Slawson, 1968).

Method

Subjects. We recruited 38 subjects from the University of Chicago community. All subjects were native speakers of American English and reported that they had no history of hearing or speech disorders and no extensive experience with synthetic speech. We paid subjects for their participation and debriefed them when they completed the experiment.

Stimuli. The stimuli consisted of the following set of nineteen monosyllabic words: "ball," "tile," "cave," "done" (the targets), "dime," "cling," "priest," "lash," "romp," "knife," "reek," "depth," "park," "gnash," "greet," "jaw," "jolt," "bluff," and "cad" (the distractors).[1] One token of each stimulus was produced by each of four synthetic talkers. We constructed the first talker (*male-150*) from one of the standard sets of parameters that are built into the DECtalk synthesizer: a "male" with an average F0 of 150 Hz. The second talker (*male-160*) was identical to the first, except we raised his average F0 to 160 Hz. The third talker (*male-300*) was identical to the first, except we doubled his average F0 to 300 Hz. The fourth talker (*female-300*) was another of the standard DECtalk female voices. All of "her" vocal characteristics were quite different from the male talkers', except we set her average F0 to 300 Hz to match male-300. The DECtalk settings used for the male and female talkers are presented in the Appendix. Note that despite the male talkers all having the same DECtalk settings aside from average pitch, average pitch interacted with other DECtalk parameters to yield modest shifts in vowel space.

To create pairs of stimuli that differed only in average amplitude, we generated a second set of stimuli produced by male-150, with average root-mean-square (RMS) amplitude digitally increased by 6 dB. We refer to this talker as *male-150-loud.* Thus, we had substantial differences in amplitude (male-150 vs. male-150-loud), pitch (male-150 vs. male-300), and vowel space (male-300 vs. female-300) along with a modest difference in pitch (male-150 vs. male-160).

Procedure. A speeded target-monitoring task was used, and hit rate, false alarm rate, and response times were calculated. Subjects were presented with an orthographic form of a target word on a computer display and were instructed to press the space bar whenever they heard the target word they saw on the screen. In each trial, subjects heard a sequence of 16 words with an SOA of 750 ms. Four targets were randomly placed at ordinal positions between the 1st and 16th stimuli. On each trial, subjects listened for one target from the set "ball," "tile," "cave," and "done." The 12 distractors were randomly selected from the full set of stimuli, excluding the trial's target.

For a pair of talkers, subjects performed the monitoring task in a blocked-talker condition and a mixed-talker condition. In the blocked trials, all of the targets and distractors were produced by one talker. In the mixed trials, equal numbers of targets and distractors were produced by both talkers and randomly ordered. Each target item served as the target 12 times in both blocked- and mixed-talker conditions (so when male-150 and male-300 were the talkers, "ball" would occur a total of 24 times—12 times per talker—in blocked- and mixed-talker conditions). Thus, for any talker pair, there were 48 blocked- and 48 mixed-talker trials.

Three different groups of 10 subjects performed the experiment with stimuli produced either by male-150 and male-300 (differing mainly in pitch), male-150 and female-300 (differing in pitch and vowel space), or male-150 and male-150-loud (differing only in amplitude). We counterbalanced talker-condition order (blocked or mixed) and order of blocked-talker blocks across subjects. One group of 8 subjects performed the experiment with stimuli produced by two talker pairs: male-150 and male-160 (differing only modestly in pitch) and male-300 and female-300 (differing mainly in vowel space). For this group, we counterbalanced talker pair, stimulus-condition order, and order of blocked trials across subjects.

Results and Discussion

We tested for effects of talker condition on hit rate, false alarm rate, and response time for each talker pair. Response times were measured from stimulus onset, and responses that occurred less than 150 ms after the onset of a stimulus were scored as responses to the preceding item. We conducted ANOVAs on the effects of talker condition for each talker pair. Hit rates were greater than .97 in both talker conditions for every talker pair, and there were no reliable effects of talker condition on hit rate (Fs = 0.16–1.60, ps = .70–.24).

False alarm rates varied from .01 to .03, but the effect of talker condition was not significant for four of the five pairs (Fs = 0.05–1.20, ps = .82–.30). There was a borderline effect for the male-150–male-160 pair, $F(1, 7) = 5.51$, $p = .051$, $\omega^2 = .01$, but in the opposite direction than would be expected (blocked = .01, SD = .01; mixed = 0, SD = 0). We examined individual differences and found that although no subjects made a false alarm response in the mixed-talker condition for this pair, in the blocked condition, 2 subjects made one false alarm, 1 made two, and 1 made five. The latter was more than 2 standard deviations from the mean. When this subject is removed, the effect on false alarms is no longer reliable, but the effects on hit rate and response time are unchanged. Therefore, we report results with this subject included.

Response times for both talker conditions for each talker pair are shown in Figure 2. The effect of talker condition was significant for three of the five pairs. A substantial change in pitch (male-150–male-300) was sufficient to slow processing reliably in the mixed-talker condition, $F(1, 9) = 11.70$, $p = .008$, $\omega^2 = .11$, as was a change primarily in vowel space (male-300–female-300), $F(1, 7) = 5.89$, $p = .046$, $\omega^2 = .08$, and changes in both pitch and vowel space (male-150–female-300), $F(1, 9) = 16.61$, $p = .003$, $\omega^2 = .15$. In contrast, neither a change in amplitude (male-150–male-150-loud) nor a small change in pitch (male-150–male-160) resulted in reliable talker-condition effects ($Fs < 1$; and the mean differences were actually slightly in the opposite of the predicted direction).

These results suggest that simple distraction is unlikely to be the underlying cause of talker variability effects, because distinct changes in amplitude did not slow processing (with the caveat that the amplitude manipulation may have been too small; however, a substantially larger amplitude manipulation is not feasible, because one talker would have to be made uncomfortably loud or too soft to allow

[1] We were constrained to using vowels for Experiment 1 because these were the materials Nusbaum and Morin (1992) recorded for their talkers. We used words when possible in the later experiments because these contain the sorts of coarticulatory cues proposed to be important under some intrinsic normalization theories.

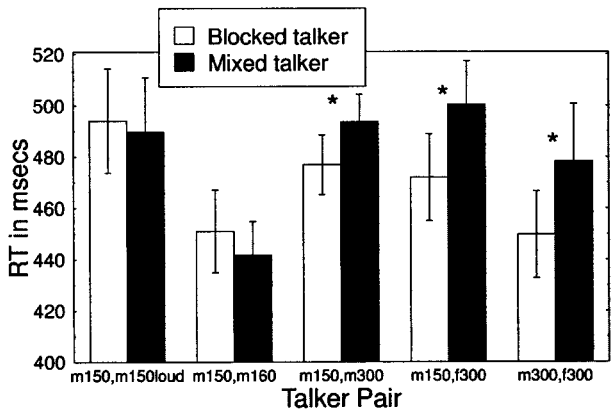

Figure 2. Response times (RTs) for each talker pair in blocked- and mixed-talker conditions in Experiment 2. Asterisks indicate statistically significant differences. Error bars represent standard error. m = male; f = female.

accurate performance). However, substantial changes in vowel space, pitch, or both were sufficient to slow processing in the mixed condition. This result provides some basic descriptive information about what sorts of acoustic changes trigger normalization.

However, the absence of an effect of the small pitch difference points to the ambiguity that remains from the absence of an effect for the two similar female talkers in Experiment 1: Did we fail to find evidence for normalization because the talkers were similar or because subjects did not detect (implicitly or explicitly) that they were different? Indeed, if one compares the differences in vowel space between the two natural female talkers in Table 1 with the differences between male-150 and male-160 in Table 2, the two pairs of talkers appear to be approximately equally similar (leaving aside timbre and other differences between the female talkers). To resolve the ambiguity that follows from the absence of an effect for some talker pairs in Experiment 2, we designed Experiment 3 to examine the distinctiveness of the similar talkers in this experiment and the similar natural female talkers from Experiment 1.

Experiment 3

We have discussed two possible explanations for the lack of mixed-talker effects for the similar natural female talkers from Experiment 1 and the synthetic male talkers differing by 10 hz in F0 from Experiment 2. First, it may be that normalization is not required for discriminable but highly similar talkers, because the same vowel space mapping fits both. Or it may be that the talkers were not discriminable, and therefore, no effect of mixing them was found—that is, talker changes were not detected, and therefore, normalization was not triggered, or indexical processing was not carried out, or both. We designed Experiment 3 to distinguish between these explanations. We asked subjects to listen to pairs of stimuli that were either identical tokens or examples of the same utterance produced by two different speakers. We asked subjects to judge either whether the two stimuli were identical tokens or whether they were produced by the same talker. We expected that if subjects were able to distinguish the similar talker pairs, they would be able to achieve high talker discrimination accuracy.

Method

Subjects. We recruited 26 subjects from the University of Chicago community. All subjects were native speakers of American English and reported that they had no history of hearing or speech disorders and no extensive experience with synthetic speech. We paid subjects for their participation and debriefed them on completion of the experiment.

Stimuli. The stimuli were produced by four pairs of talkers: three synthetic pairs from Experiment 2 (male-150–male-300, male-300–female-300, and male-150–male-160) and the pair of natural female talkers from Experiment 1 (natural female 1 and natural female 2). For the synthetic talkers, the stimuli consisted of the word list used in Experiment 2. For the natural talkers, the stimuli consisted of the eight isolated vowel stimuli from Experiment 1 augmented with a set of eight /rVk/ syllables that had been recorded previously for other purposes ("rack," "reek," "rick," "rock," "rook," "ruck," "ruke," and "wreck").

Procedure. On each trial, subjects heard a pair of examples of the same vowel or word. Both items were either produced by the same talker or by different talkers. For one group of subjects (the *talker* group), the instructions were to press a key marked *different* only if the two stimuli sounded as if they were produced by different talkers and to press a key marked *same* if the talkers sounded like the same person. For the other group of subjects (the *token* group), the instructions were to press a key marked *different* if the two stimuli sounded different in any way and to press *same* if both stimuli were the same recording. Both groups of subjects were told to respond as quickly and accurately as possible.

For each talker pair, there were 40 trials on which the expected response was *same*, with the same stimulus repeated by the same talker; half of these used speech from one talker, and half used speech from the other. The expected response in both the talker and token conditions was *same* for these pairs of repeated items. There were also 40 trials on which the two stimuli were produced by different talkers, and the first stimulus was produced by each talker on half of the trials. The expected response in both task conditions was *different* for these 40 pairs.

Results and Discussion

We conducted separate one-way ANOVAs on the talker-pair factor for same and different trials for the talker and token groups.

Table 2
Vowel Spaces (in Hz) of the Synthetic Talkers Used in Experiments 2–4

		Vowel					
Talker	Formant	/i/	/I/	/a/	/U/	/u/	/æ/
Male 150	F1	340	490	744	540	410	645
	F2	2,120	1,745	1,165	1,110	1,010	1,660
Male 160	F1	360	460	785	515	440	660
	F2	2,070	1,755	1,195	1,040	1,040	1,635
Male 300	F1	380	480	660	580	350	645
	F2	1,995	1,755	1,005	1,040	995	1,610
Female 300	F1	315	430	740	470	370	920
	F2	2,800	2,020	1,215	1,280	1,395	2,030

Two subjects were excluded from the talker group because they had extremely low accuracy (<.50) for different trials for the dissimilar talkers, suggesting they were not performing the tasks correctly. In general, accuracy was high, and far above chance, for all conditions, indicating that even the small pitch and/or vowel space differences between the pairs of similar talkers were sufficient to lead to the perception of distinct talkers.

Accuracy and response latencies for the token group are shown in the left panels of Figure 3. The effect of talker pair was not significant for same token accuracy, $F(3, 33) = 1.74$, $p = .179$; $\omega^2 = .004$, with mean accuracy greater than .99 for every talker pair. Nor was the effect significant for same response time, $F(3, 33) < 1$. However, there was a reliable effect of different token accuracy, $F(3, 33) = 5.33$, $p = .004$, $\omega^2 = .03$, with subjects substantially less accurate on the similar talker pairs (around .90) than on the dissimilar pairs (around .99). Despite a trend toward longer latencies for the similar pairs, the effect was not reliable for response time, $F(3, 33) = 1.89$, $p = .15$, $\omega^2 = .007$.

The results were very similar for the talker conditions (see the right panels of Figure 3). The effect of talker pair was not significant for accuracy or response time for same-talker trials ($Fs < 1$). However, the effect was reliable for different-talker trials, $F(3, 27) = 6.11$, $p = .003$, $\omega^2 = .05$, with substantially lower accuracy for the similar talkers (.85–.91) than for the dissimilar talkers (.95–.97). The effect of response time was also reliable, $F(3, 27) = 7.67$, $p = .001$, $\omega^2 = .07$, with substantially longer response times for the similar talkers.

The results are not unequivocal. On the one hand, subjects had significantly greater difficulty distinguishing the similar talkers, suggesting that the absence of mixed-talker effects in Experiments 1 and 2 for the similar pairs could have been due to subjects failing to detect talker differences. On the other hand, performance was still far above chance, and it would seem implausible that a 6%–10% decrease in the rate of *different* responses to the similar talkers versus the dissimilar talkers would reflect a degree of confusability that would eliminate mixed-talker effects.

However, the task demands in this experiment were quite different from those in the monitoring tasks. Subjects were explicitly asked to say whether tokens or talkers were different. This likely influenced them to focus on discovering possible differences. The 6%–10% decrease in accuracy for similar talkers might mask a much higher level of confusability in the absence of instructions to explicitly look for differences. Furthermore, individual accuracy for different token and talker judgments for the similar talkers ranged from .53 to 1.0. This suggests that the degree of subjective difference varied from subject to subject (though median accuracy was .93), and the absence of mixed-talker effects could stem from a subset of subjects who failed to detect differences between talkers. However, explicit awareness of a talker change is not required for normalization to occur in contextual tuning theory—rather, the hypothesize–test–adjust mechanism triggers evaluation of talker characteristics when the current mapping cannot provide a good fit between the input and phonemic categories (or lexical items, etc.).

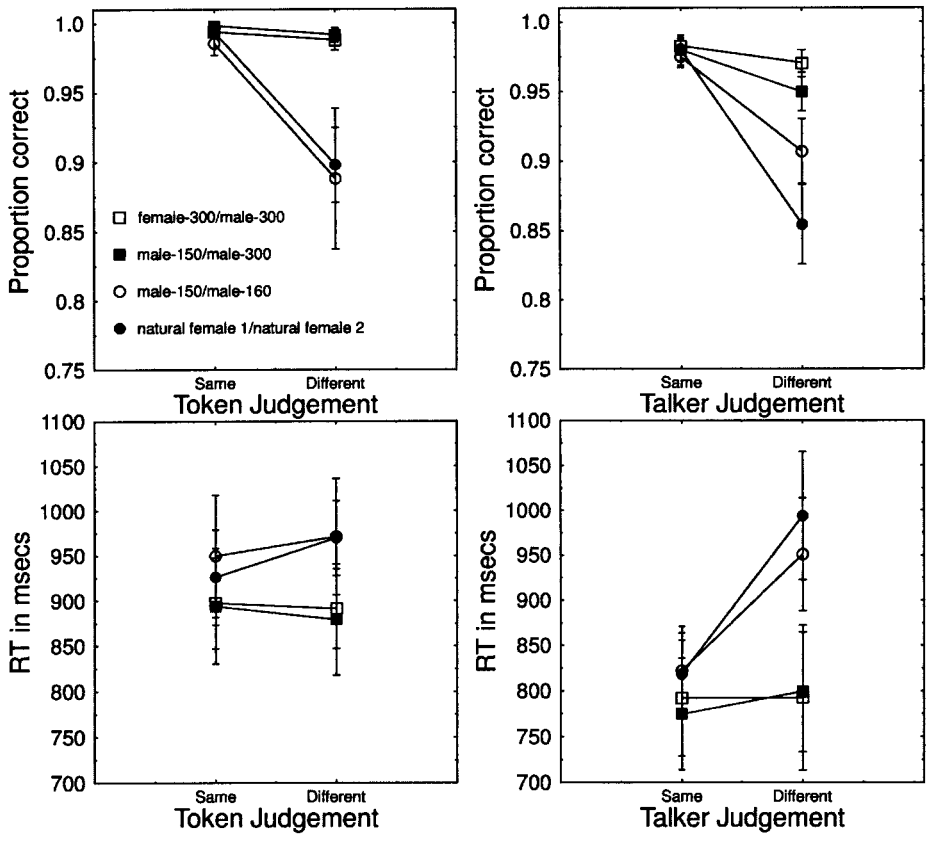

Figure 3. Accuracy and response times (RTs) for token and talker judgments for each talker pair in Experiment 3. Error bars represent standard error.

What, then, determines the effects of talker variability? In Experiment 3, we directed listeners to respond to acoustic differences that could signal a difference between talkers. In our previous monitoring tasks, listeners were directed to recognize words regardless of talker differences. Thus, it is possible that listeners may not treat a small but discriminable acoustic difference between talkers as a change in talker unless they have nonphonetic (or even nonlinguistic) reasons to do so. We carried out a fourth experiment to test this possibility.

Experiment 4

Nusbaum and Magnuson (1997) argued that linguistic processes and speech perception in particular are active processes that are influenced by expectations derived from context. The psycholinguistic literature is rife with examples of context dependence. Among the examples in spoken language are the word superiority effect (i.e., that phonemes can be detected more quickly in words than nonwords; Rubin, Turvey, & van Gelder, 1976), phoneme restoration (context-dependent restoration of a phoneme replaced with noise or an ambiguous sound as a function of lexical or sentential context; e.g., Samuel, 1981, 1997; Warren, 1970), and immediate integration of visual context to constrain sentence processing (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). However, proponents of autonomous models of language processing (e.g., Norris, McQueen, & Cutler, 2000), in which there is strict encapsulation of information at distinct stages of processing such that the system is completely feedforward, argue that such results reflect postperceptual decisions. Although there are other demonstrations that lexical knowledge influences phonemic processing in a top-down, online fashion (Elman & McClelland, 1988; Magnuson, McMurray, Tanenhaus, & Aslin, 2003a; Samuel, 2001; Samuel & Pitt, 2003), these results remain controversial (see the discussion between McQueen, 2003, and Magnuson, McMurray, Tanenhaus, & Aslin, 2003b, and between McClelland, Mirman, & Holt, 2006, and McQueen, Norris, & Cutler, 2006).

A truly compelling demonstration of an active system would show effects of nonlinguistic, cognitive expectations on putatively low-level aspects of speech perception. Perhaps the best-known and most dramatic demonstration is the finding of Remez, Rubin, Pisoni, and Carrell (1981) that subjects' expectations strongly influenced their perceptions of a three-tone sinusoidal replica of a naturally produced sentence. When they were given no information about the stimulus, subjects were more likely to perceive it as a variety of nonspeech sounds (e.g., "science fiction" or bird sounds) than as speech. When they were told to expect computer-generated speech, however, more than half the subjects accurately transcribed the entire utterance. This is evidence that perceptual processes are directed to different aspects of a source signal when subjects expect to hear speech, but it may reflect the inducement of a mode of processing rather than a continuous modulation of perception on the basis of expectations.

Carden, Levitt, Jusczyk, and Walley (1981) found that consonant distinctions based on place of articulation can be influenced by expectations for manner of articulation, but this still runs afoul of the autonomous, postperceptual criteria—as does any study that depends on direct measures of phonetic decisions, because one cannot disentangle perceptual and decision processes in such results (cf. Samuel, 2001). Closer to our topic, K. M. Fenn, Atkins, Skipper, Bond, and Nusbaum (2004) found that when subjects being interviewed by telephone were told to expect a talker change after being put on hold, they reported having heard two talkers even when the talker did not change. This suggests that the processing of talker differences could be susceptible to expectations.

Could the significance of acoustic variability correlated with talker variability also depend on expectations? Given our hypothesis that talker processing depends on active control (given, e.g., the Nusbaum & Morin [1992] finding of load sensitivity), one might expect that when there is a borderline difference between talkers, the mechanisms sensitive to talker differences might be susceptible to top-down cognitive influence. That is, although awareness of a talker change is not a prerequisite for normalization, the mere expectation that a talker change has occurred may be sufficient to trigger normalization.

We tested this possibility in Experiment 4 by asking two groups of subjects to do the monitoring task used in Experiment 2 with the male-150–male-160 pair for which we did not find a talker condition effect. However, we explicitly told subjects whether to expect to hear one talker or two. If mixed-talker effects reflect mandatory processing of indexical characteristics or a passive, automatized normalization process that is part of an autonomous language processing system, cognitive expectations should not be able to change them.

Method

Subjects. We recruited 16 subjects from the University of Chicago community. All subjects were native speakers of American English and reported that they had no history of hearing or speech disorders and no extensive experience with synthetic speech. We paid subjects for their participation and debriefed them on completion of the experiment.

Stimuli. The stimuli consisted of the 19 monosyllabic words produced by talkers male-150 and male-160 used in Experiment 2.

Procedure. The monitoring paradigm described for Experiment 2 was used, and all blockwise parameters were maintained with the materials produced by male-150 and male-160. Each target was presented 12 times in the blocked condition (in which all targets and distractors were produced by the same talker) and 12 times in the mixed condition (in which the talker changed randomly from word to word).

Subjects were randomly assigned to two groups of 8. The only difference between groups was the instructions they received prior to the experiment. The *1-voice expectation* group was told that they would be hearing synthetic speech produced by a single synthetic talker; to make the speech sound more natural, we sometimes changed the pitch, but there was only one talker. Subjects heard a monologue with some pitch variation (with average F0 jumping from 150 to 160 hz). The *2-voice expectation* group was told that we had created two synthetic talkers by modifying pitch and that in some parts of the experiment, they would hear speech from two talkers. They also heard a dialogue between the two voices. The monologue and dialogue were both approximately 30 s long. The instructions are shown in Figure 4.

Results and Discussion

As in Experiments 1 and 2, we measured accuracy and response times. Response times were measured from stimulus onset, and responses that occurred less than 150 ms after the onset of a stimulus were scored as

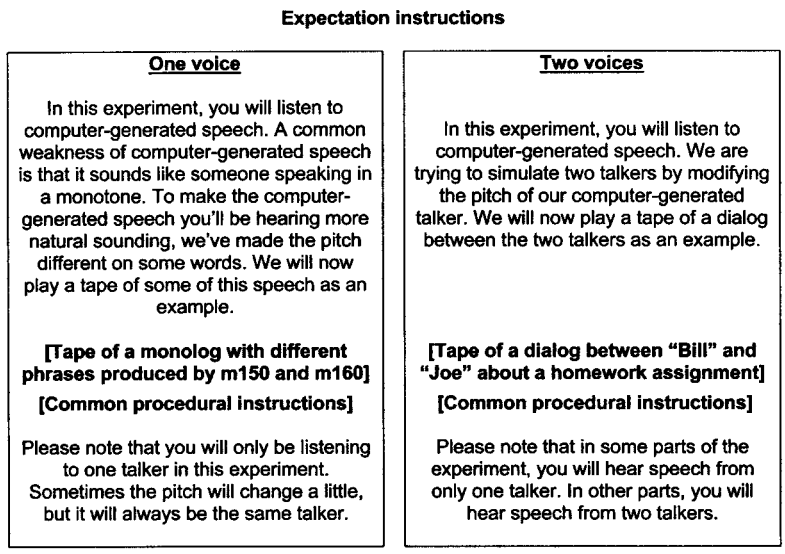

Figure 4. The one- and two-voice expectation instructions for Experiment 4.

responses to the preceding item. There were no reliable differences between performance in blocked and mixed trials for the 1-voice expectation group in hit rate (blocked = .98, SD = .02; mixed = .98, SD = .03); false alarm rate (blocked = .02, SD = .02; mixed = .03, SD = .03), $F(1, 7) = 2.1$, $p = .20$, $\omega^2 = .01$; or response time (blocked = 430 ms, SD = 82; mixed = 430 ms, SD = 77). For the 2-voice expectation group, however, subjects were reliably faster to respond in blocked trials (M = 470 ms, SD = 55) than in mixed trials (M = 492 ms, SD = 51), $F(1, 7) = 6.2$, $p = .042$, $\omega^2 = .09$. Under 2-voice instructions, there were no reliable differences in hit rate (blocked = .97, SD = .04; mixed = .97, SD = .04) or false alarm rate (both means and standard deviations were .02). Figure 5 shows the response time results from the two expectation conditions and the results from Experiment 2 for this talker pair for comparison. Figure 6 presents individual response time data for the 1- and 2-voice expectation groups and the neutral instruction group. Note that most subjects in the neutral and 1-voice instruction groups show no effect of talker condition or trend toward slower processing. A striking reversal is observed in the 2-voice instruction group, in which all but 2 subjects are substantially slower in the mixed condition than in the blocked condition.

The only difference between the two groups was the expectation developed by the preexperiment instructions—the stimuli were identical. This indicates that the talker effects we have observed are not the result of a purely passive process (e.g., template matching, as van Bergem, Pols, & Koopmans-van Beinum [1988] suggest, or simple exemplar-based resonance; Goldinger, 1998). Rather, phonetic processing is constrained by knowledge of nonacoustic states in the world—whether these take the form of expectations about the source of the acoustic signal (e.g., the number of talkers, knowledge that the signal is intended as speech [Remez et al., 1981]), visual information (e.g., regarding articulations [McGurk & McDonald, 1976] or communicative gestures [McNeill, 1992, e.g., p. 134]). This strongly suggests that phonetic constancy in speech perception relies on active control structures.[2]

General Discussion

What constitutes a difference between talkers? How do listeners detect (implicitly or explicitly) a change in talker? Clearly, one way is to perceive a sufficiently large acoustic difference between talkers. When talkers differ (primarily) either in vowel space (as in Experiment 1) or fundamental frequency (as in Experiment 2), the difference between talkers slows down processing. This talker variability effect is reliable and consistent with previous results using the same paradigm (Nusbaum & Morin, 1992) and with studies using different methods (e.g., Mullennix & Pisoni, 1990). However, this talker variability effect is not simply a result of acoustic variability. First, not all discriminable talker differences result in a talker variability effect. In Experiment 1, the two female talkers were quite different in voice quality, but their vowel spaces were sufficiently similar that no reliable talker variability effect was observed when the talkers were mixed. Second, discriminable variation in amplitude does not slow processing. This indicates that

[2] An anonymous reviewer suggested an alternative interpretation of the basis for the effect. Rather than alerting listeners to the presence of two talkers, the expectation instructions may have directed subjects' attention to the pitch differences during the dialogue more effectively than did the single-talker instructions (despite containing specific mention of the use of pitch differences as an attempt to make the talker sound more natural). We find this explanation unlikely and less parsimonious than one that ties this result to typical processing of talker differences. There is also reason to expect that perceptual learning that accompanies exposure to a talker involves rather different processes than those involved in adjusting to a talker change. K. F. Fenn, Nusbaum, and Small (2004) examined the cortical activity accompanying perceptual learning of a synthetic talker. Their results suggest that perceptual learning of a talker reduces cortical activity in the posterior part of left superior temporal gyrus (STG). However, when confronted with talker variability, Wong et al. (2004) showed that activity in the STG increases. Moreover, during training, learning seems to recruit the motor system, increasing activity in ventral premotor, but talker variability seems to increase activity in the superior parietal region. Thus, talker variability and talker learning seem to have different effects on cortical processing. Our next step will be to examine whether the neural correlates of talker-condition effects based on expectations resemble those for genuine differences, using a design based on that of Wong et al.

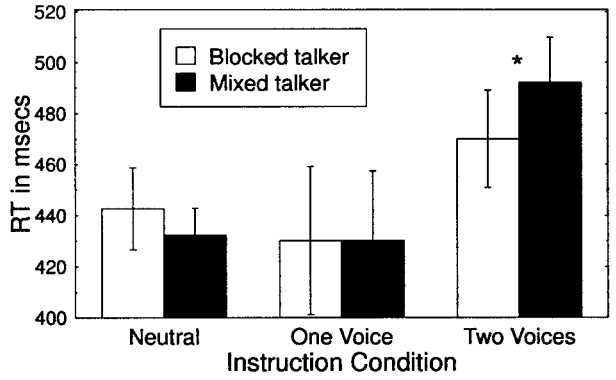

Figure 5. Response times (RTs) for the one- and two-voice conditions in Experiment 4 along with the results for the same talker pair in Experiment 2 (neutral instructions). The asterisk indicates a statistically significant difference. Error bars represent standard error.

simple discriminable variation in acoustic quality is not sufficient to produce talker variability effects.

When does talker-specific acoustic variability slow speech recognition? First, when the vowel spaces are sufficiently different between talkers, as in Experiment 1, recognition is significantly slower. This is entirely consistent with the talker normalization view that listeners must interpret acoustic cues in the context of the vocal characteristics of the talker. It could also be construed as consistent with mandatory encoding of indexical vocal characteristics of the new talker (Nygaard & Pisoni, 1998), although such a talker encoding perspective might suggest that for any discriminable talker difference, recognition should slow, which does not happen. Experiment 1 also showed that when the pitch difference is sufficiently large between talkers, recognition slows in the mixed-talker condition. This, too, is consistent with the view that listeners simultaneously encode talker vocal characteristics and phonetic information (the indexical–phonetic contingency view of Nygaard & Pisoni, 1998). It is also consistent with a talker normalization view, given the assumption that large pitch differences between talkers would be ecologically associated with differences in vowel spaces. Our listeners (and their perceptual systems) did not know that we had dissociated fundamental frequency and vocal tract size using a computer speech synthesizer. However, it is the third case (i.e., Experiment 4) that is the most interesting. Given a constant, modest pitch difference between two synthetic talkers with identical vowel spaces, listeners' expectations determine whether or not there is a variability effect when the speech of these talkers is mixed. If the pitch difference is treated as within-talker variability, mixing two sets of utterances differing in pitch does not slow processing, whereas if the pitch difference is *expected* to correspond to between-talker variability, the mixed condition slows processing.

On the one hand, this suggests that talker identity has the same many-to-many, lack-of-invariance problem that holds for phonetic segments. The same acoustic difference between speech samples could arise because of one speaker using different speaking styles or two talkers differing in that characteristic. And a real difference between talkers might be manifested with little or no acoustic difference. The problem of adjusting to changes in talker characteristics then might be thought of as the same kind of computational problem as recognizing phonetic structure (cf. Nusbaum & Magnuson, 1997). In other words, detecting talker differences that require perceptual accomodation is itself a perceptual problem that may not be handled automatically or passively.

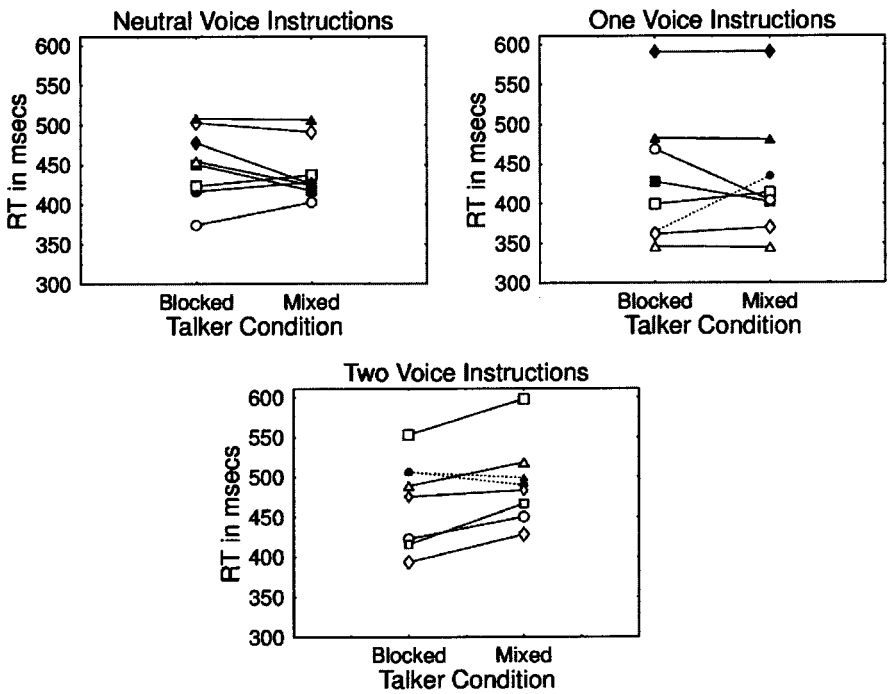

Figure 6. Individual response times (RTs) for the one- and two-voice conditions in Experiment 4, along with the results for the same talker pair in Experiment 2 (neutral instructions).

On the other hand, the cognitive penetrability of talker difference detection raises questions about some aspects of the standard view of speech perception. What is the "standard view"? Given the disparities that exist between theories such as the motor theory of speech perception at one extreme (Liberman & Mattingly, 1985) and auditory theories at the other (e.g., Diehl et al., 2004), it would seem difficult to find any aspect of processing theories that is standard. However, even across this range of types of theories, there is a strong flavor of modularity (cf. Fodor, 1983), whether stated explicitly (as in Liberman & Mattingly, 1985) or not. These theories take speech perception as fast, automatic, encapsulated, and independent of other cognitive processes.

The present finding that expectations about the interpretation of an acoustic signal changes the nature of the processing of that signal fits well within a series of studies in speech perception that should be troublesome to any kind of modular, purely bottom-up, passive processing theory. Remez et al. (1981) showed that simply telling listeners that sine wave signals are speech was enough to shift processing from an apparently nonlinguistic auditory mode to linguistic interpretation. Carden et al. (1981) showed that when listeners "imagined" (following instructions) one manner of articulation not present in a consonant series, perception of the place of articulation shifted as if appropriate cues to manner were acoustically present in the signal. Niedzielski (1999) showed that telling listeners that a set of synthetic vowels were produced by a speaker from Windsor, Ontario, Canada, shifted vowel classification relative to telling the listeners the speaker came from Detroit. Johnson, Strand, and D'Imperio (1999) showed pictures of male and female faces paired with "androgynous" male and female speech and found expectations based on the pictures changed perception of the speech.

In all these studies, high-level cognitive information (e.g., beliefs, expectations) changed the interpretation of an acoustic signal as if different signal properties were present. The experience in these studies is as salient perceptually as is the McGurk effect (McGurk & Macdonald, 1976)—this is not some kind of logico-deductive process wherein listeners are supplying compliant answers. In our Experiment 4, the effects of expectation were a 25-ms increase in response time in the mixed condition, which is squarely in the 20–30-ms increase that is found with genuinely distinct talkers. Together, these studies raise the same question originally framed by Remez et al. (1981) as to whether the role of traditional acoustic speech cues needs revision.

Rather than view speech perception as mediated by simple passive mapping systems that relate cues to categories, researchers need to think about speech perception in a fundamentally different way. We have previously argued (Nusbaum & Magnuson, 1997) that almost all theories of speech perception treat the relationship between acoustic cues and linguistic categories as a deterministic mapping, assuming that one has only to discover the information in or transformations of the speech signal that will provide invariance. An invariant mapping would allow speech perception to be explained via passive computations using open-loop control structures (Nusbaum & Magnuson, 1997; Nusbaum & Schwab, 1986). However, given the frustrating failure to discover a basis for invariance, it may be necessary to accept that the mapping between acoustic cues and linguistic categories is truly nondeterministic. If one acoustic pattern can be classified as a member of multiple categories depending on context, this is formally a nondeterministic system, and the mapping cannot be resolved by the kinds of theories that have been proposed previously.

However, there is an alternative theory that has been argued to provide an account of phonetic constancy without appeal to active control structures: nonanalytic episodic theories of speech perception and spoken word recognition (Goldinger, 1998; Nygaard & Pisoni, 1998), in which holistic (unanalyzed) episodic traces of events provide the basis for processing speech. Both proposals subscribe to the principle of *lawful variability* (Elman & McClelland, 1988)—that is, that variability in the speech signal is informative about the sources of variability and, therefore, is information rather than noise to be abstracted away (Nygaard, 2005). Because the nonanalytic account promises to explain phonetic constancy without active control, it bears discussion here.

Nonanalytic Episodic Theories

Nonanalytic theories are motivated by two sets of results. First, there is a perceptual contingency between recognizing phonetic and indexical properties of speech. Mullennix and Pisoni (1990) found that listeners cannot ignore irrelevant variation in indexical characteristics (e.g., talker identity) when their task is to attend to phonetic information, and vice versa (although the relationship is asymmetrical; talker variability interfered more with phonetic processing than vice versa, which Mullennix & Pisoni, 1990, interpreted as indicating that phonetic processing is hierarchically contingent [cf. Turvey, 1973] on assessment of talker characteristics, because talker characteristics condition phonetic realization, but the converse is not equally true), suggesting the two sorts of information are processed integrally (Garner, 1974). Consistent with this result, Nygaard and colleagues (Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994) reported that training on talker identification promotes better perception of speech produced by the trained-on talkers.

Pisoni and his colleagues (e.g., Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994; Pisoni, 1997) have argued that the interaction between talker vocal characteristics and the acoustic–phonetic structure of speech does not require any kind of perceptual compensation or normalization process.[3] Instead, they argued that when there is a change in talker, listeners attend to both talker vocal characteristics and phonetic information, encoding both into long-term memory and, more radically, that this happens in a *nonanalytic* fashion. Both indexical and phonetic characteristics of speech must be attended to simultaneously because they are inseparable; the speech signal is not

[3] An additional motivation given for nonanalytic episodic accounts is the claim that any sort of normalization implies throwing out nonlinguistic information, which would be inconsistent with evidence for memory and processing effects of surface details of speech. Although this view has been part of some accounts of normalization, normalization itself does not entail throwing out information. For example, there is nothing in Joos's (1948) proposals that indicate that information is discarded. Instead, the proposal is that a coordinate mapping is achieved between the characteristics of the talker one is hearing and internal categories. However, evidence for sensitivity to surface detail would not preclude even a normalization account on which the mapping is achieved by extracting only linguistically relevant cues from speech—if it is assumed that multiple representations of the sensory input are maintained, and at least one contains surface details. This would be a reasonable position given the substantial evidence for multiple representations at different levels of abstraction in vision (e.g., Milner & Goodale, 1995), audition (Bushara et al., 1999), and speech (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000).

decomposed but is processed holistically. From this perspective, the demands of attending to both sets of properties are the putative cause of the adverse effects of talker variability.

The second phenomenon motivating nonanalytic episodic accounts is that memory for and processing of spoken words is affected by "nonlinguistic" talker characteristics. Performance on old–new list memory tasks is better when words are repeated in the same voice rather than different voices (Church & Schacter, 1994; Goldinger, 1996; Goldinger, Pisoni, & Logan, 1991; Palmeri, Goldinger, & Pisoni, 1993). Thus, even after words are recognized, there is some interaction between the linguistic and indexical properties of an utterance, at least in terms of retrieving episodic traces of the words.

Goldinger (1998) proposed an episodic theory of lexical access to account for these phenomena. In seeking to model memory effects of surface details, Goldinger proposed that speech is perceived via memory traces of word-length episodes that preserve all surface details, and the speech signal is neither isolated nor decomposed. The episodic approach is a "kitchen-sink" version of lawful variability that approximates a deterministic mapping from nondeterministic input with a passive control structure. By including all aspects of the signal (in particular, indexical properties and sub- and supraphonemic signal variation), a statistical compositing mechanism (based, in Goldinger's [1998] implemented model, on MINERVA2; Hintzman, 1988) is theorized to be able to create clusters of episodes that will allow context independent speech perception and word recognition, without explicit normalization.

How would the episodic model do this? Consider the traditional view of invariance (reflected both by Lane [1965] and by Liberman et al. [1967]), that each phoneme *could* be realized as a unique (or a collection of unique) physical patterns, but the process of speech production obscures and restructures these patterns. Coarticulatory influences of proximal phonetic segments, physical and other differences in the vocal characteristics of talkers, and differences in speaking rate all are sources of acoustic perturbations that impose noise on the true phonetic signal pattern. Note the similarity of this view of linguistic classification to the statistical problem of detecting, in any kind of experiment, the influence of an independent variable on a measured dependent variable (cf. Anderson, Silverstein, Ritz, & Jones, 1977) given all the sample error and measurement noise that occurs in real world measurements. This metaphor for pattern classification as a statistical memory-retrieval process underlies the mathematics of the MINERVA2 model (Hintzman, 1988), which is the foundation of the implementation of Goldinger's (1998) episodic lexicon model.

The statistical compositing process by which retrieved representations are aggregated from experienced auditory representations of utterances is quite similar to the idea of the law of large numbers in statistics in discerning the mean of a distribution from increased sampling of a population. At least in the abstract, this notion of statistical recognition serves the same purpose as normalization mechanisms in other theoretical approaches. In the Goldinger (1998) model (based on MINERVA2; Hintzman, 1988), the statistical aggregation of episodic traces of auditory experiences of speech similar to an utterance abstracts a linguistic form even without such abstractions being stored explicitly (see McClelland & Rumelhart, 1985). Thus, irrelevant surface acoustic detail seems to be "filtered" out, and apparently abstract phonetic forms are recognized even though the model does not explicitly represent such abstract forms in memory.

Goldinger's (1998) model was proposed explicitly as an alternative to normalization theories. The idea of normalization is that sources of acoustic variability need to be taken into account to determine context-specific acoustic–phonetic mappings. Goldinger (1998) represented normalization for talker variability as filtering the variability away from the phonetic pattern to be recognized, and related nonanalytic theories regard evidence for sensitivity to subphonemic and nonlinguistic surface detail as direct evidence against normalization (e.g., Pisoni, 1997). These arguments assume a view of normalization on which abstract phonetic representations are stored mentally, and for a listener to match their corresponding auditory patterns to any particular utterance, perturbations of those patterns by variation in phonetic context, talker differences, and speaking rate must be filtered out. This corresponds to a particularly old view of normalization in the computer vision literature in which an input pattern must be normalized into a canonical view orientation, position, and size before matching against stored object models (Roberts, 1965). However, this is quite different from the view of normalization proposed by Joos (1948) or that found in most subsequent normalization theories of speech perception (for further discussion, see Nusbaum & Magnuson, 1997, and Footnote 3).

The nonanalytic episodic model potentially accounts for all but one of the talker variability phenomena we have reviewed. It proposes that the invariance problem can be solved through richer representations. It could account for processing declines in mixed-talker conditions via resonance: Because speed of recognition depends on the strength of the response when a new trace is compared with traces in memory, and more recent traces are represented more strongly, a performance advantage should emerge in the blocked-talker condition because of accumulating recency benefits (see Magnuson et al., 2007, for a comparison of short- and long-term effects of talker familiarity, and further discussion of the theories under review here). It is not clear that the model can account for context effects such as those reported by Ladefoged and Broadbent (1957)[4] or for evidence for active control (attentional modulation and interactions with working memory [Nusbaum & Morin, 1992] and the expectation-based variability effects in Experiment 4). However, it naturally accounts for effects of linguistic and nonlinguistic talker characteristics on memory and processing and for the contingency of those characteristics (e.g., Goldinger, 1996, 1998; Nygaard & Pisoni, 1998).

However, if we turn to the larger question of phonetic constancy across all kinds of variability, serious questions arise for nonanalytic episodic theories. The first problem is that the model assumes segmentation; words are the basic episodic unit. The second problem is category formation. Goldinger (1998) argued that his model obviates normalization: A passive, exemplar-matching process would not need

[4] It is possible that effects of recency might also account for this phenomenon. In the Ladefoged and Broadbent (1957) experiment, the relevant stimulus was a sample without cues to nonphonetic talker characteristics and was thus ambiguous between two phonemic categories. If it is assumed that the carrier sentence activates the characteristics of one talker, that may be sufficient to disambiguate the consonant–vowel–consonant words that must be classified. However, note that "talker characteristics" in this study consisted simply of vowel spaces, and that these had to be inferred, in part, over multiple samples of speech; it appears that accounting for this initial step would be challenging for the episodic theory.

special mechanisms to accommodate talker differences—it would simply activate the appropriate cluster on the basis of trace similarity. This, however, begs the question of how categories would form—what would be the basis for clustering two talkers' acoustically distinct vowels in a single category? This point has not yet received detailed treatment in the theory, but it exactly parallels the normalization question: What are the stimulus properties that are used to classify speech samples as exemplars of phonemic categories? An additional problem is that the simulations Goldinger (1998) presented in support of the claim that normalization is not required depend on an assumption that differed importantly from the larger nonanalytic assumption of the theory: Indexical and phonetic features were represented as independent vectors in those simulations. This implicitly assumes analysis to separate the features, because in real speech, talker characteristics condition the realization of phonetic information; they cannot be had for free in an unanalyzed episode.

Our true difference with nonanalytic episodic theories (Goldinger, 1998; Pisoni, 1997) is not with the exemplar-based approach but with the denial of analysis. The statistical compositing we have just described might go far toward providing at least a partial solution to acoustic–phonetic mapping—if it were made analytic. Assuming the input is literally nonparametric (raw, unanalyzed wholes) begs crucial questions (segmentation, cluster formation, etc.). Analysis is also required to account for the evidence for active control we have already reviewed as well as a growing body of results documenting the amazing adaptability of speech and word perception and production. For example, Dell, Reed, Adams, and Meyer (2000) and Onishi, Chambers, and Fisher (2002) have shown, in production and perception, respectively, that speakers and listeners are exquisitely sensitive to short-term changes in phonotactic probabilities, and Maye, Aslin, and Tanenhaus (2003) found that listeners adapt very quickly to novel vowel space rotations. In the latter case, highly learned acoustic–phonetic mappings must be suppressed to successfully perceive speech from the rotated vowel space. How could attention modulate the weight of any parameter of speech if the input is a parameter-free, unanalyzed episode? What is needed is an analytic approach that includes active control.

Active Control Mechanisms

We have suggested that the solution to nondeterminism in speech perception may lie in control structures—specifically, active, closed-loop mechanisms capable of adapting quickly to changes in acoustic–phonetic mappings (cf. Nusbaum & Magnuson, 1997). Understanding phonetic constancy may require us to consider more cognitive, active processes wherein multiple hypotheses regarding the interpretation of an acoustic pattern are tested so as to recognize speech (e.g., on the basis of the success of lexical access or integration with discourse context; cf. Newell, 1975). Such hypothesis tests might necessitate shifts in attention to such different cues in the speech signal as would be appropriate to the specific context. For example, Nusbaum and Morin (1992) reported that when there is talker variability, listeners use different acoustic cues than they do when there is a single talker. More recently, Wong et al. (2004) used fMRI imaging to show that talker variability increases attentional processing in traditional speech areas (e.g., posterior superior temporal cortex) but also recruits areas that are more associated with shifts of attention (i.e., superior parietal cortex).

Note that neither the claim that contextual tuning invokes attentional processing nor an appeal to the notion of hypothesis testing entails that the listener be explicitly aware of the operation of the underlying mechanisms, that the listener must explicitly detect a talker change, or even that a talker change must be implicitly detected to trigger normalization. This is an important point to consider, given findings that listeners sometimes fail to detect talker changes (K. M. Fenn, Atkins, et al., 2004; Vitevitch, 2003) and that false experiences of talker change can be induced (K. M. Fenn, Atkins, et al., 2004).[5]

An active, hypothesis-testing theory of speech perception (e.g., Stevens & Halle, 1967) can use multiple sources of information to constrain proposed hypotheses and to test those hypotheses. Given the nondeterministic mapping present in speech, parallel tests of alternative possible interpretations provide a plausible basis that could ultimately resolve the lack-of-invariance problem. As we discussed earlier, evidence for active control is consistent with a variety of architectures that provide approximate-test-adjust mechanisms. These range from interactive-activation models (McClelland & Elman, 1985; McClelland & Rumelhart, 1981); to simple recurrent networks (Elman, 1990), and attractor networks (Plaut & Kello, 1999), in which recurrent connections allow nondeterministic mappings to be approximated; to conflict-monitoring networks of the sort developed by Cohen and his colleagues (Botvinick, Braver, Carter, Barch, & Cohen, 2001; Cohen, Braver, & O'Reilly, 1996). Although these models have focused on how highly overlearned responses can be moderated under task demands (as in the Stroop task), they also provide a potential framework for attentional modulation more generally.

What of the episodic effects of (putatively) linguistically irrelevant details of speech events? Another possibility would be an analytic exemplar-based approach compatible with active control. Johnson (1994, 1997) proposed an analytic exemplar model in

[5] It is crucial to note the distinction between evidence for change deafness and blindness and an assertion that the changed aspects of a stimulus have not been processed. Change blindness or change deafness occur in situations in which the environment is typically stable, and change is unexpected. Change blindness should not imply a failure to process, pre- and postchange, the aspects of a scene that change, nor even a complete failure to process the change itself—there is instead a failure of awareness. For example, Hayhoe, Bensinger, and Ballard (1998) found increased fixation durations following scene changes in a simple copying task that subjects did not report noticing, but only when the change was task-relevant (see also Triesch, Ballard, Hayhoe, & Sullivan, 2003). Similarly, Luck, Vogel, and Shapiro (1996) examined event-related potentials (ERPs) to words presented during attentional blinks. Subjects could not recall the presented word, but online ERP patterns indicated that the words were processed semantically, because their meanings influenced N400 magnitude. Wolfe (1999) argued that such findings show that change blindness and inattentional blindness (Mack & Rock, 1998) would be better described as inattentional *amnesia*. The literature on visual search is instructive on this point. Despite some correspondence between subjective and empirical measures of attentional allocation in search under automatized processing (typically construed as parallel search) and controlled processing (typically construed as serial search), subjects have at best little awareness of the underlying mechanisms or of the gradual changes in processing that occur as a task is automatized (see also Wolfe, 1998, for challenges to the parallel–serial division, and a review of evidence instead for a continuum of difficulty). In short, a process under active control is not the same thing as a process under strategic control.

which normalization depends on attentional modulation of parameters (in the generalized context model framework; Nosofsky, 1988), ranging from specific acoustic features, such as F0, to indexical cues or, potentially, any relevant information, such as visual cues indicating a talker change. This framework is compatible with all the talker variability phenomena we have reviewed, including evidence for active control.

However, the active control mechanism is an externally applied contextual attentional modulation and, so, is itself unspecified. Goldinger and Azuma (2003) presented a very similar approach, in which they abandon nonanalysis. The new theoretical framework they sketched is explicitly analytic (indeed, references central to the earlier arguments for nonanalysis [e.g., Jacoby & Brooks, 1984] were no longer cited, and the word *nonanalytic* did not appear). They proposed that adaptive resonance theory (Grossberg, 1980) may provide a unifying framework in which a hierarchical structure of features at multiple scales (e.g., acoustic, phonetic, phonemic, syllabic, words) is built up through experiences of contingent activation. Similar structures form for different sorts of sensory data and knowledge, and sufficient contingent activation across modalities or knowledge types could result in perceptually coherent structures that will "resonate" (form part of a set of recurrently connected structures) and allow for, for example, social influences on phonetic perception (Goldinger & Azuma, 2003). However, again, attention remains an external modulator rather than an integral part of the system. A possibility for future exploration would be to use mechanisms like those in Cohen and colleagues' conflict monitoring networks (Botvinick, Braver, Carter, Barch, & Cohen, 2001; Cohen et al., 1996), which allow attentional control without homunculuslike external modulation.

Lawful Variability as the Guiding Principle

The common factor among recent alternative approaches to nondeterminism in speech perception, and talker variability in particular (Nusbaum & Magnuson, 1997; Nygaard, 2005), is the appeal to Elman and McClelland's (1988) observation that variability in speech is not noise. Rather, variability reveals lawful contingencies relevant for phonetic perception as well as other types of knowledge carried in speech (e.g., talker identity, emotional state). But how can the wheat be told from the chaff—which aspects of variability are relevant for phonetic constancy? Nonanalytic episodic theories are conservative to a logical extreme: An omnimodality episode preserves everything. This approach provides an explanation for episodic effects on perception and memory as well as state-dependent learning and other effects of implicit memory. Without an existence proof that such a system can handle a realistic analogue of variability in speech (or speech itself), this approach provides little insight into the informational structure of speech or the cognitive, perceptual, and neural mechanisms that underlie phonetic constancy. The true insight of the nonanalytic episodic approach is the need to account for exemplar effects and the possibility that something like an omnimodal memory trace may play a role in some aspects of speech processing.

We have argued that a nonanalytic mechanism could not account for attentional, working-memory, and expectation-based results that indicate active control structures underlie speech processing. Active control structures also provide a potential computational solution to the lack-of-invariance problem. Evidence for active control is compatible with a variety of approaches, including analytic exemplar models that allow active control (Goldinger & Azuma, 2003; Johnson, 1994, 1997), attractor networks (e.g., Plaut & Kello, 1999), or any of these mechanisms coupled with an active control mechanism, with one candidate being the conflict monitoring framework (Botvinick et al., 2001; Cohen et al., 1996).

But still missing are general principles that might provide insight into the contingencies in speech that are relevant for phonetic constancy. Features are stipulated rather than discovered in all of the computational models we have discussed. Work on the *triangle model* of visual word recognition (e.g., Harm & Seidenberg, 2004) may provide a heuristic approach for speech. In the triangle model, a minimal set of representational nodes are posited: one for visual features (orthography), one for phonology, and one for semantics. These are minimal in that they are the smallest set needed to account for the interfaces to written language. Pathways exist from each set to the others in an attractor network framework. As the network is trained (e.g., to produce the correct phonological pattern given a particular orthographic input), contingencies are discovered and embedded in the network through connection-weight learning. The model becomes sensitive to the relative lawfulness of the orthography–phonology mapping compared with the largely arbitrary semantic mappings. Plaut and Kello (1999) proposed a model in the triangle framework for speech perception and production, using the most realistic analogues to speech used in any current model. The contingencies such a model is able to learn are those that are predictive of one or more mappings. Modeling may provide one route to discovering the contingencies (lawful variability) relevant for phonetic constancy by heuristically adding sets of interface nodes and using the model to discover contingencies between interfaces.

Conclusion

The heavy lifting of identifying specific mechanisms remains, but significant progress in understanding phonetic constancy could be gained by considering the role of control structures in resolving nondeterminancy. However, regardless of the correct account of phonetic constancy, the present experiments demonstrate that under certain conditions, listeners' expectations are as powerful as acoustic cues in triggering processes that allow perceptual accommodation of talker variability.

References

Ainsworth, W. (1975). Intrinsic and extrinsic factors in vowel judgments. In G. Fant & M. Tatham (Eds.), *Auditory analysis and perception of speech* (pp. 103–113). London: Academic Press.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review, 84,* 413–451.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000, January 20). Voice-selective areas in human auditory cortex. *Nature, 403,* 309–312.

Best, C. T. (1994). The emergence of native language phonological influences in infants: A perceptual assimilation model. In J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167–224). Cambridge, MA: MIT Press.

Botvinick, M. M., Braver, T. S., Carter, C. S., Barch, D. M., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108,* 624–652.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker,

rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics, 61,* 206–219.

Bushara, K. O., Weeks, R. A., Ishii, K., Catalan, M. J., Tian, B., Rauschecker, J. P., & Hallett, M. (1999). Modality-specific frontal and parietal areas for auditory and visual spatial localization in humans. *Nature Neuroscience, 2,* 759–766.

Carden, G., Levitt, A., Jusczyk, P. W., & Walley, A. (1981). Evidence for phonetic processing of cues to place of articulation: Perceived manner affects perceived place. *Perception & Psychophysics, 29,* 26–36.

Church, B. A., & Schacter, D. L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 521–533.

Cohen, J. D., Braver, T. S., & O'Reilly, R. C. (1996). A computational approach to prefrontal cortex, cognitive control, and schizophrenia: Recent developments and current challenges. *Philosophical Transactions of the Royal Society of London, Series B, 351,* 1515–1527.

Creelman, C. D. (1957). Case of the unknown talker. *Journal of the Acoustical Society of America, 29,* 655.

Dell, G. S., Reed, K. D., Adams, D. R., & Meyer, A. S. (2000). Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1355–1367.

Dellatre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America, 27,* 769–773.

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

Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. (1977). Stop consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception & Psychophysics, 22,* 109–122.

Elman, J. L. (1990). Finding structure in time. *Cognitive Sciences, 14,* 179–211.

Elman, J., & McClelland, J. (1988). Exploiting lawful variability in the speech weave. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 360–380). Hillsdale, NJ: Erlbaum.

Fenn, K. F., Nusbaum, H. C., & Small, S. L. (2004, May). *Cortical mechanisms of perceptual learning of spoken language.* Paper presented at the 11th Annual Meeting of the Cognitive Neuroscience Society, San Francisco, CA.

Fenn, K. M., Atkins, A. S., Skipper, J. I., Bond, V. C., & Nusbaum, H. C. (2004). *When less is heard than meets the ear: Change deafness in a telephone conversation.* Poster session presented at the 2004 meeting of the Psychonomic Society.

Fodor, J. A. (1983). *The modularity of mind.* Cambridge, MA: MIT Press.

Fowler, C. A. (1989). Real objects of speech perception: A commentary on Diehl and Kluender. *Ecological Psychology, 1,* 145–160.

Fujisaki, H., & Kawashima, T. (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics, AU-16,* 73–77.

Garner, W. R. (1974). *The processing of information and structure.* Potomac, MD: Erlbaum.

Gerstman, L. J. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics, AU-16,* 78–80.

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,* 305–320.

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 152–162.

Gordon, P. C. (1988). Induction of rate-dependent processing by coarse-grained aspects of speech. *Perception & Psychophysics, 43,* 137–146.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review, 87,* 1–51.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review, 111,* 662–720.

Hayhoe, M. M., Bensinger, D. G., & Ballard, D. H. (1998). Task constraints in visual working memory. *Vision Research, 38,* 125–137.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528–551.

Jacoby, L. L., & Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception and concept learning. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 18, pp. 1–47). New York: Academic Press.

Johnson, K. (1994). Memory for vowel exemplars. *Journal of the Acoustical Society of America, 95,* 2977.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–166). San Diego, CA: Academic Press.

Johnson, K., Ladefoged, P., & Lindau, M. (1993). Individual differences in vowel production. *Journal of the Acoustical Society of America, 94,* 701–714.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics, 27,* 359–384.

Joos, M. (1948). *Acoustic phonetics.* Baltimore: Linguistic Society of America.

Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., & Thulborn, K. R. (1996, October 4). Brain activation modulated by sentence comprehension. *Science, 274,* 114–116.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America, 73,* 322–335.

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

Lane, H. (1965). The motor theory of speech perception: A critical review. *Psychological Review, 72,* 275–309.

Liberman, A. M. (1986). Brief comments on invariance in phonetic perception. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability of speech processes* (pp. 490–492). Hillsdale, NJ: Erlbaum.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74,* 431–461.

Liberman, A. M., Delarrte, P. C., & Cooper, F. S. (1952). The role of selected stimulus variables in perception of unvoiced stop consonants. *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.

Liu, T., Slotnick, S. D., Serences, J. T., & Yantis, S. (2003). Cortical mechanisms of feature-based attentional control. *Cerebral Cortex, 13,* 1334–1343.

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). *Journal of the Acoustical Society of America, 102,* 1134–1140.

Luck, S. J., Vogel, E. K., & Shapiro, K. L. (1996, October 17). Word meanings can be accessed but not reported during the attentional blink. *Nature, 382,* 616–618.

Mack, A., & Rock, I. (1998). *Inattentional blindness.* Cambridge, MA: MIT Press.

Magnuson, J., McMurray, B., Tanenhaus, M., & Aslin, R. (2003a). Lexical

effects on compensation for coarticulation: The ghost of Christmash past. *Cognitive Science, 27,* 285–298.

Magnuson, J., McMurray, B., Tanenhaus, M., & Aslin, R. (2003b). Lexical effects on compensation for coarticulation: A tale of two systems? *Cognitive Science, 27,* 801–805.

Magnuson, J. S., Nusbaum, H. C., & Akahane-Yamada, R. (2007). *Talker familiarity and talker normalization.* Manuscript submitted for publication.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics, 28,* 407–412.

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 676–684.

Maye, J., Aslin, R., & Tanenhaus, M. (2003). In search of the Weckud Wetch: Online adaptation to speaker accent. *Proceedings of the CUNY Conference on Sentence Processing.* Cambridge, MA.

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences, 10,* 363–369.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88,* 375–407.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114,* 159–188.

McGurk, H., & McDonald, J. (1976, December 23–30). Hearing lips and seeing voices. *Nature, 264,* 746–748.

McNeill, D. (1992). *Hand and mind.* Chicago: University of Chicago Press.

McQueen, J. M. (2003). The ghost of Christmas future: Didn't Scrooge learn to be good? Commentary on Magnuson, McMurray, Tanenhaus, & Aslin (2003). *Cognitive Science, 27,* 795–799.

McQueen, J. M., Norris, D., & Cutler, A. (2006). Are there really interactive processes in speech perception? *Trends in Cognitive Sciences, 10,* 533.

Miller, J. L., & Baer, T. (1983). Some effects of speaking rate on the production of /b/ and /w/. *Journal of the Acoustical Society of America, 73,* 1751–1755.

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

Miller, R. L. (1953). Auditory tests with synthetic vowels. *Journal of the Acoustical Society of America, 18,* 114–121.

Milner, A. D., & Goodale, M. A. (1995). *The visual brain in action.* Oxford, England: Oxford University Press.

Mullennix, J. W., & Howe, J. N. (1999). Selective attention in perceptual adjustments to voice. *Perceptual and Motor Skills, 89,* 447–457.

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47,* 379–390.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85,* 365–378.

Navon, D. (1984). Resources—a theoretical soup stone? *Psychological Review, 91,* 216–234.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America, 85,* 2088–2113.

Newell, A. (1975). A tutorial on speech understanding systems. In R. D. Reddy (Ed.), *Speech recognition* (pp. 4–54). New York: Academic Press.

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology, 18,* 62–85.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23,* 299–370.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 700–708.

Nusbaum, H. C., & Magnuson, J. S. (1997). Talker normalization: Phonetic constancy as a cognitive process. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 109–132). San Diego, CA: Academic Press.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Tohkura, Y. Sagisaka, & E. Vatikiotis-Bateson (Eds.), *Speech perception, speech production, and linguistic structure* (pp. 113–134). Tokyo: OHM.

Nusbaum, H. C., & Schwab, E. C. (1986). The role of attention and active processing in speech perception. In E. C. Schwab & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines: Vol. 1. Speech perception* (pp. 113–157). San Diego, CA: Academic Press.

Nygaard, L. C. (2005). Perceptual integration of linguistic and nonlinguistic properties of speech. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 390–413). Malden, MA: Blackwell.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60,* 355–376.

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

Onishi, K. H., Chambers, K. E., & Fisher, C. (2002). Learning phonotactic constraints from brief auditory experience. *Cognition, 83,* B13–B23.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 309–328.

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

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

Plaut, D. C., & Kello, C. T. (1999). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In B. MacWhinney (Ed.), *The emergence of language* (pp. 381–415). Mahwah, NJ: Erlbaum.

Posner, M. I. (2003). Imaging a science of mind. *Trends in Cognitive Sciences, 7,* 450–453.

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 40–61.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981, May 22). Speech perception without traditional speech cues. *Science, 212,* 947–950.

Roberts, L. G. (1965). Machine perception of three-dimensional solids. In J. T. Tippett (Ed.), *Optical and electro-optical information processing.* Cambridge, MA: MIT Press.

Rubin, P., Turvey, M. T., & van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in spoken nonwords. *Perception & Psychophysics, 19,* 384–398.

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General, 110,* 474–494.

Samuel, A. G. (1997). Lexical activation produces potent phonemic percepts. *Cognitive Psychology, 32,* 97–127.

Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science, 12,* 348–351.

Samuel, A. G., & Pitt, M. A. (2003). Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory and Language, 48,* 416–434.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84,* 1–66.

Shankweiler, D., Strange, W., & Verbrugge, R. (1977). Speech and the problem of perceptual constancy. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing* (pp. 315–345). Hillsdale, NJ: Erlbaum.

Slawson, A. W. (1968). Vowel quality and musical timbre as functions of

spectrum envelope and fundamental frequency. *Journal of the Acoustical Society of America, 43,* 87–101.

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition: I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America, 96,* 1314–1324.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America, 64,* 1358–1368.

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

Stevens, K. N., & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. In W. Walthen-Dunn (Ed.), *Models of the perception of speech and visual form* (pp. 88–102). Cambridge, MA: MIT Press.

Strange, W. (1989). Dynamic specification of coarticulated vowels spoken in sentence context. *Journal of the Acoustical Society of America, 85,* 2135–2153.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America, 79,* 1086–1100.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M. & Sedivy, J. E. (1995, June 16). Integration of visual and linguistic information in spoken language comprehension. *Science, 268,* 1632–1634.

Triesch, J., Ballard, D. H., Hayhoe, M. M., & Sullivan, B. T. (2003). What you see is what you need. *Journal of Vision, 3,* 86–94.

Turvey, M. (1973). On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli. *Psychological Review, 80,* 1–52.

van Bergem, D. R., Pols, L. C., & Koopmans-van Beinum, F. J. (1988). Perceptual normalization of the vowels of a man and a child in various contexts. *Speech Communication, 7,* 1–20.

Vitevitch, M. S. (2003). Change deafness: The inability to detect changes in a talker's voice. *Journal of Experimental Psychology: Human Perception and Performance, 29,* 333–342.

Warren, R. M. (1970, January 23). Perceptual restoration of missing speech sounds. *Science, 167,* 392–393.

Wolfe, J. (1998). Visual search. In H. Pashler (Ed.), *Attention* (pp. 13–74). Philadelphia: Taylor & Francis.

Wolfe, J. (1999). Inattentional amnesia. In V. Coltheart (Ed.), *Fleeting memories: Cognition of brief visual stimuli* (pp. 71–94). Cambridge, MA: MIT Press.

Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience, 16,* 1173–1184.

Appendix

DECtalker Parameters for Three Synthetic Talkers

Parameter	Talker	
	Male-150/160/300	Female-300
Sex	1	0
Head size	100	95
F4	3300	4500
F5	3650	2500[a]
F4 bandwidth	260	230
F5 bandwidth	330	2048
Average F0	150/160/300	300
F0 range	100	135
Voicing source gain	65	65
Aspiration source gain	70	70
Frication source gain	80	84
Nasalization gain	74	70
Aascade formant resonator 1 gain	68	67
Cascade formant resonator 2 gain	60	65
Cascade formant resonator 3 gain	48	51
Cascade formant resonator 4 gain	64	58
Loudness of voice[b]	86	80
Breathiness	0	0
Lax breathiness	0	50
Smoothness	3	60
Richness	70	100
Fixed samples of open glottalization (*n*)	0	10
Laryngealization (%)	0	0
Baseline fall	18	8
Hat rise	18	20
Stress rise	32	32
Assertiveness	100	100
Quickness	40	30

Note. All of the male talkers used in the study differed only in average F0.
[a] Setting F4 lower than F5 effectively turns off higher formants in DECtalk. [b] For the experiments, root-mean-square (RMS) amplitude was digitally normalized.

Received October 22, 2005
Revision received April 13, 2006
Accepted April 19, 2006 ■