Research Article

First Impressions and Last Resorts

How Listeners Adjust to Speaker Variability

Tanya Kraljic,[1] Arthur G. Samuel,[2] and Susan E. Brennan[2]

[1]University of California, San Diego, and [2]Stony Brook University

ABSTRACT—*Perceptual theories must explain how perceivers extract meaningful information from a continuously variable physical signal. In the case of speech, the puzzle is that little reliable acoustic invariance seems to exist. We tested the hypothesis that speech-perception processes recover invariants not about the signal, but rather about the source that produced the signal. Findings from two manipulations suggest that the system learns those properties of speech that result from idiosyncratic characteristics of the speaker; the same properties are not learned when they can be attributed to incidental factors. We also found evidence for how the system determines what is characteristic: In the absence of other information about the speaker, the system relies on episodic order, representing those properties present during early experience as characteristic of the speaker. This "first-impressions" bias can be overridden, however, when variation is an incidental consequence of a temporary state (a pen in the speaker's mouth), rather than characteristic of the speaker.*

The paradox of perception is that even though the physical properties of stimuli vary continuously, people's perceptions of those stimuli are remarkably constant. People recognize faces they have never viewed twice from exactly the same angle in exactly the same lighting, and they identify words from phonetic segments that vary substantially from word to word and from speaker to speaker. The great mystery of perception is how such constancy is achieved. Within the field of speech perception, the dominant approach to investigating this question has focused on the information available in the speech signal, the assumption being that there must be acoustic invariants in the signal that are extracted during perception (e.g., a constant acoustic pattern that always accompanies the percept of an [s]). However, researchers have been unable to find a set of invariants that work for all contexts and speakers. Recent research has examined alternative mechanisms, including *perceptual learning*: Rather than searching for acoustic invariants (that may not exist), listeners sometimes restructure their phonemic category boundaries to accommodate the variation they hear (see, e.g., Kraljic, Brennan, & Samuel, in press; Kraljic & Samuel, 2005, 2006; Maye, Aslin, & Tanenhaus, in press; Norris, McQueen, & Cutler, 2003).

In some sense, restructuring the underlying representations seems at odds with the idea of having stable representations that give rise to perceptual constancy. What does it mean to say that listeners perceive [s] if the properties that define an [s] change in response to the input? A balance between flexibility and stability is critical to all cognitive and perceptual systems, and requires that learning cannot proceed indiscriminately; not every acoustic variation that is encountered should be learned. Some variations are incidental or transient: They do not reflect how that speech sound is normally pronounced by that speaker. Suppose a friend has had too much to drink. Your perceptual system should not "learn" that your friend slurs his words or devoices final consonants; the next time you encounter him, you would then have to learn that he does not. Although the perceptual system must be flexible enough to adjust to variations in pronunciation, it would not be desirable to reset the underlying representations every time there is a temporary change of state.

We propose that learning is restricted to *characteristic* pronunciations. This proposal diverges profoundly from dominant accounts of speech perception, as information about whether a variant is characteristic is not "in" the acoustic signal. Yet recent evidence (Kraljic et al., in press) suggests that listeners distinguish different sources of variation even when the acoustic realization is identical. For example, we looked at perceptual learning of a specific acoustic-phonetic variation of [s], in which

Address correspondence to Tanya Kraljic, UCSD Center for Research in Language (CRL), 9500 Gilman Dr., La Jolla, CA 92093-0526, e-mail: tkraljic@crl.ucsd.edu.

[s] was pronounced midway between [s], and [ʃ] ("sh"), a pronunciation we refer to as [~sʃ]. We found that listeners who heard this pronunciation learned it when it reflected a stable characteristic of a speaker (when all [s] segments were pronounced as [~sʃ]), but not when it was specific to a particular phonetic context (as in dialects that realize [s] as [~sʃ] only in [str] contexts).

The idea that the perceptual system extracts invariants not of the proximal stimuli themselves, but of the sources underlying those stimuli, features prominently in Gibson's (1966, 1973) ecological theory of perception, although the theory is vague about how this is accomplished. According to articulation-based theories of speech perception (e.g., Fowler, 1986; Liberman & Mattingly, 1985), information about a variation's source is available directly from articulatory gestures. This idea, too, is consistent with our proposal that pronunciations resulting from stable speaker characteristics (e.g., devoiced consonants due to a foreign accent) will be learned, but those that are incidental (devoiced consonants due to intoxication) will not. However, the articulation-based approach simply reformulates the question, rather than providing a solution. How could a listener perceive (or recover) different gestures for the same acoustic realization under different pragmatic circumstances? If the perceptual system is to learn in a more discriminating fashion, it must use other information to determine what to learn.

We suggest that the system could use pragmatic information to assess variation, and thereby to guide perception and learning. Our hypotheses were derived by analogy to the *generic-viewpoint assumption* from visual perception (Binford, 1981; Freeman, 1994; Rock, 1983): In the absence of evidence to the contrary, the perceptual system presumes that the sample of information it is processing is neither accidental nor improbable. This assumption appears to underlie many feats of visual disambiguation (e.g., recovering edges and objects; Albert & Hoffman, 2000). Because spoken language unfolds over time, the parallel assumption in the auditory domain logically results in a bias toward interpreting initial events as stable. Properties present during the first encounter are assumed to be characteristic and should therefore be learned.

In fact, computational approaches support just such a "first-impressions" bias: When unconstrained, a neural network often settles on a solution or representation too quickly, so that the solution is based heavily on initial exposure to some input. The network is subsequently unable to recover when the solution turns out to be incorrect (a problem often referred to as the problem of local error minima; e.g., A. Clark & Thornton, 1997; Elman, 1993). In the study reported here, our first manipulation tested whether such a primacy bias constrains perceptual learning when adults hear noncanonical variations in pronunciation.

Our second manipulation made the leap between learning and pragmatics: If relying on order reflects an attempt to constrain the space of hypotheses about the input in order to learn the correct "solution" as quickly as possible, then strong pragmatic evidence about whether the initial input is actually likely to be representative of future input (e.g., "this speaker is visibly drunk, so slurring is not a permanent attribute of his speech") should override such a reliance.

In sum, we predicted that phonetic space is restructured only when a particular pronunciation is attributable to an enduring characteristic of the speaker, and that the perceptual system uses two kinds of extralinguistic information to make this attribution (and therefore to determine whether a pronunciation will be learned): First, in the absence of any explicit cues to permanence, it will rely on a first-impressions bias, and second, it will use pragmatic cues when they are available and strong.

We used an established perceptual-learning paradigm (Norris et al., 2003) to test our hypotheses. Listeners first heard a speaker whose pronunciation of a particular segment was ambiguous between [s] and [ʃ] (i.e., [~sʃ]). For half of the listeners, the ambiguous [~sʃ] sound replaced [s] in 10 words (such as *episode*); for the other half, [~sʃ] replaced [ʃ] (e.g., in *vacation*). In each condition, the 10 critical words were randomly interspersed among many other words and nonwords. Listeners identified each item as a word or a nonword. If exposure to [~sʃ] resulted in perceptual learning, listeners' subsequent perception of the critical sound (either [s] or [ʃ]) should have shifted to include the [~sʃ] variant.

To test for such learning, we asked all participants to complete a category-identification test immediately following exposure. Listeners heard items ranging on a continuum from [s] to [ʃ] and pressed a button to indicate (for each item) whether they heard "S" or "SH." Perceptual learning would be indicated if listeners who heard [~sʃ] embedded in [ʃ]-words categorized more items as [ʃ], and if listeners who heard [~sʃ] embedded in [s]-words categorized fewer items as [ʃ].

To examine when the system engages perceptual learning, we manipulated two exposure factors: modality of presentation (audio only vs. audiovisual) and attribution for the odd pronunciation (characteristic of the speaker vs. incidental consequence of some temporary state). The audiovisual modality provided explicit evidence about whether a pronunciation was characteristic or incidental, whereas the audio-only modality required relying more heavily on inferences.

We predicted that in the absence of any other information about the speaker, the perceptual system is biased toward first impressions: Properties present in a person's speech when he or she is first encountered are learned as characteristic for that speaker. Accordingly, the attribution factor for the audio-only modality was an order manipulation: During exposure, listeners heard both normal and ambiguous pronunciations of the critical phonetic segment from the same speaker; for example, listeners heard both *epi?ode* (with [s] replaced by [~sʃ]) and *parasite* (with unambiguous [s]). Critically, for half of our listeners, ambiguous tokens preceded normal ones (characteristic condition); for the other half, ambiguous tokens followed normal ones (incidental condition). Our hypothesis predicted that listeners

hearing ambiguous pronunciations first would attribute them to the speaker and show learning; listeners hearing them second would instead attribute them to some (unknown) transient cause and not show learning.

Our manipulation of attribution in the audiovisual modality provided an additional test. In this modality, ambiguous pronunciations were always presented before normal ones. Note that in the audio-only case, we predicted perceptual learning under these conditions. Crucially, for half the audiovisual subjects, we simultaneously provided a reason for the odd pronunciation that was external to the speaker (incidental condition): She had a pen in her mouth when she produced the ambiguous tokens. Our hypothesis (that the system is designed to learn things that are permanent about the speaker) predicted that subjects who saw the pen would not show perceptual learning. Those who heard the same pronunciations, in the same order, but with no suggested reason for the ambiguity (characteristic condition), were expected to show perceptual learning.

METHOD

Participants

Two hundred sixty-eight students participated for research credit. All were 18 years of age or older and identified themselves as native English speakers with normal hearing. The 128 participants in the audio-only conditions were students at Stony Brook University; the 140 participants in the audiovisual conditions were students at the University of California, San Diego.

Materials

Phase 1—Exposure (Lexical Decision)

For the auditory lexical-decision task, we created two experimental lists, each with 100 words and 100 nonwords. The lists were identical except that they differed in which 10 words were pronounced with [∼s∫].

Stimulus Selection. The stimuli are described in detail elsewhere (Kraljic & Samuel, 2005). The 100 words in each list included 20 words that contained [s] (but no [∫]), 20 words that contained [∫] (but no [s]), and 60 filler words that contained no [s] or [∫]. The 40 [s]- and [∫]-words ranged in length from two to four syllables. Each [s] or [∫] occurred in the initial position of a syllable that was relatively late in the word, to ensure that the critical phoneme was preceded by enough of the word to generate strong lexical activation. The two sets of critical words ([s]- and [∫]-words) were matched in mean syllable length and frequency. The 60 filler words matched the critical words in stress pattern, number of syllables, and word frequency.

To ensure equal numbers of "word" and "nonword" responses during the lexical-decision task, we created 100 filler nonwords with no [s] or [∫]. Each participant thus heard 100 words and 100 nonwords: In the ?S condition, there were 10 normally pronounced [s]-words, 10 [s]-words in which [s] was replaced with [∼s∫], 20 normal [∫]-words, 60 filler words, and 100 filler nonwords; in the ?SH condition, the [s]-words were all normally pronounced, and instead 10 of the [∫]-words contained [∼s∫].

Stimulus Construction. All of the words and nonwords were recorded by a female speaker, who also produced a second version of each [s]- and [∫]-word in which she replaced the critical phoneme that normally appeared in the word with the other phoneme (e.g., *episode* and *epishode*). The word pairs were used to create an ambiguous [∼s∫] mixture for each critical word. The acoustic properties of [s] and [∫] allow a relatively straightforward mixing of the waveforms to construct [∼s∫] ([s] and [∫] are similar in both duration and amplitude). The [s]-version of each word was used as the "frame." The [s] and [∫] were mixed together with five weightings that varied from 30% [s] and 70% [∫] to 70% [s] and 30% [∫] (see Kraljic & Samuel, 2005, for details). Attempting to keep the degree of ambiguity constant across the stimuli, we selected a single ambiguous mixture for each word to use in the experiment.

Audio-Only Conditions. Our central hypothesis was that the perceptual system takes the drastic step of shifting perceptual boundaries only when it does not have a way to account for odd input. In the audio-only conditions, we manipulated the attribution factor by varying when the odd input was encountered. For one group of participants, the 10 tokens containing [∼s∫] were randomly interspersed among the first 100 lexical-decision items, and the 10 normal versions of the critical sound occurred in the last 100 items; for the other half of the participants, the 10 normally produced versions were in the first half of the list, and the 10 odd versions were in the second half. Thus, four stimulus-presentation lists were created. The lists differed in whether or not the odd pronunciation was present during the first half of the exposure phase (normal first vs. odd first) and in whether the critical ambiguous pronunciation replaced [s] or [∫] (?S vs. ?SH). We expected that if listeners heard the normal versions first, they would attribute the odd pronunciations that followed to some (unknown) transient articulation problem. In contrast, when the odd pronunciations were not preceded by normal versions from that speaker, there was no alternative attribution suggested by the stimuli, and the system would undergo perceptual learning.

Audiovisual Conditions. For all participants in the audiovisual conditions, the odd tokens occurred within the first half of the list. This is the order that should produce perceptual learning. However, for half of the audiovisual participants, the stimuli offered an alternative attribution for the odd pronunciations, potentially allowing the perceptual system to avoid restructuring. For all audiovisual participants, video information was presented simultaneously with the audio information. In the video, the speaker fidgeted with a pen, and on half of the trials, she put the pen in her mouth as she spoke. In one condition (incidental

pronunciation), the speaker always had the pen in her mouth on critical trials (i.e., when the pronunciation of the critical sound was [∼s∫]). In the other condition (characteristic pronunciation), the speaker never had the pen in her mouth on critical trials. The audio track was taken directly from the audio-only condition and spliced onto a video of the speaker.

The audiovisual stimuli were constructed as follows: The speaker was seated against a blank backdrop, with a pen in her right hand. The audio stimuli were played to her one by one, and she repeated each word or nonword, taking care to imitate the rate of speech. Throughout the recording, she fidgeted with the pen in her hand. She recorded the entire list twice, each time putting the pen in her mouth on a different 50% of the items. The video and audio were recorded onto digital videotape and edited using Adobe Premier®.

We saved each item as a separate .avi file, including 15 frames of silence both before and after the word (or nonword) was spoken. We replaced the first 5 of these silent frames with a blank (black) screen, which was followed by 10 frames during which the screen faded into the speaker. The 16th frame marked the onset of the word (or nonword). After the word's offset, we performed the fade-out to black video over the final 15 frames. Finally, for each item, we replaced the audio that was produced during filming with the audio for that same item in the audio-only condition.

Four stimulus-presentation lists were created. The lists differed in whether or not the speaker had the pen in her mouth during the 10 critical items with the ambiguous pronunciation (incidental vs. characteristic pronunciation) and in whether the critical ambiguous pronunciation replaced [s] or [∫] (?S vs. ?SH).

Phase II—Category Identification

In the second phase of the experiment, all participants heard six items on a continuum that ranged from [asi] to [a∫i], spoken in the same voice as the lexical-decision items. The procedure for creating the continuum was similar to that for creating the ambiguous critical items used in the lexical-decision task: Each endpoint ([asi] and [a∫i]) was recorded, and the [s] and [∫] were mixed together in proportions varying from 20% [s] and 80% [∫] to the reverse. Six mixtures that ranged from relatively [s]-like to relatively [∫]-like were chosen. Presentation was strictly auditory for all participants.

Procedure

Participants were randomly assigned to one of the eight possible lists for the lexical-decision task. In the audio-only conditions, up to 3 participants were tested simultaneously in a sound-attenuated booth. In the audiovisual conditions, participants were tested individually on a laptop computer in a quiet room. For all participants, the audio stimuli were presented over headphones, and participants responded "word" or "nonword" by pressing the corresponding button on a response panel; responses and reaction times were recorded. The instructions emphasized both speed and accuracy. Participants were not told that some of the items might have ambiguous sounds.

After the lexical-decision phase, all participants categorized sounds on the [asi]-[a∫i] continuum, with each of the six items presented 10 times. One participant was replaced because of extremely low (20%) accuracy on the lexical-decision task.

RESULTS

Lexical Decision

The lexical-decision data measure how quickly and accurately participants responded to words pronounced with an ambiguous fricative ([∼s∫]), compared with words in which the fricative was pronounced naturally. Listeners performed very well overall (see Table 1). Across the four exposure conditions (combinations of modality and attribution), mean accuracy was 97.6% for the naturally pronounced items and 94.9% for the items with the ambiguous fricative. In the two audio-only exposure conditions, accuracy was slightly but significantly higher for the natural versions of the critical items than for the ambiguous versions—characteristic pronunciation: $t_1(126) = 5.23$, $p_{rep} = .99$, and $t_2(58) = 2.88$, $p_{rep} = .97$; incidental pronunciation: $t_1(126) = 4.09$, $p_{rep} = .99$, and $t_2(58) = 5.72$, $p_{rep} = .99$. In neither of the

TABLE 1

Mean Accuracy and Reaction Times (for Correct Items) on the Lexical-Decision Task

	Pronunciation of critical words			
	Natural		Ambiguous [∼s∫]	
Condition	Accuracy (%)	Reaction time (ms)	Accuracy (%)	Reaction time (ms)
Audio-only				
No alternative attribution (ambiguous tokens first)	98.5 (0.05)	1,077 (104)	93.1 (0.6)	1,079 (171)
Alternative attribution (ambiguous tokens last)	99.5 (0.01)	1,045 (161)	96.1 (0.4)	1,030 (111)
Audiovisual				
No alternative attribution (no pen in mouth)	96.7 (0.07)	1,133 (151)	95.3 (0.4)	1,516 (173)
Alternative attribution (pen in mouth)	95.5 (0.19)	1,130 (237)	95.1 (0.8)	1,163 (217)

Note. Standard deviations are given in parentheses.

audiovisual exposure conditions was the slight accuracy difference between natural and ambiguous items reliable. Similarly, for three of the four exposure conditions, there were no reliable differences in reaction times between the natural and ambiguous fricative stimuli. However, in the audiovisual, characteristic-pronunciation condition, the ambiguous stimuli did yield significantly slower responses than the naturally pronounced stimuli, $t_1(138) = 14.3$, $p_{rep} = .99$, and $t_2(57) = 11.75$, $p_{rep} = .99$. Collectively, the lexical-decision data suggest that our [∼s∫] mixtures were relatively natural sounding.

Category Identification

Overall Perceptual Learning

For each participant, we calculated the average percentage of test syllables identified as "SH." To the extent that learning had occurred, listeners exposed to [∼s∫] in [∫]-words should have identified more syllables as "SH" than listeners exposed to that sound in [s]-words. Indeed, we found a main effect of perceptual-learning condition, $F(1, 260) = 7.17$, $p_{rep} = .96$. This effect did not interact with presentation modality, $F(1, 260) < 1$, $p_{rep} = .03$. Thus, learning occurred to the same extent whether exposure was audio only or audiovisual. Critically, perceptual learning depended on our attribution manipulation, $F(1, 260) = 5.08$, $p_{rep} = .92$: In the two conditions in which listeners had to attribute the odd pronunciation to an idiosyncrasy of the speaker (characteristic), listeners learned the speaker's pronunciation. In contrast, no such learning occurred in the two conditions in which other attributions about the pronunciation could be made (incidental). Figure 1 shows the pattern of perceptual-learning effects (and noneffects) in the two modalities.

Characteristic Variations Are Learned, Incidental Variations Are Not

The pattern of results was the same in both modalities, indicating that the perceptual system distinguishes whether or not the [∼s∫] pronunciation is characteristic. Figure 2 shows the perceptual-learning effect in the audio-only modality. Listeners who heard ambiguous pronunciations first learned those pronunciations: On the categorization test, there were more "SH" responses for participants in the ?SH condition (58.9%) than for those in the ?S condition (51.8%), $F(1, 62) = 5.93$, $p_{rep} = .93$, $d = 0.6$. In contrast, participants who heard normal pronunciations first did not show perceptual learning (59.7% "SH" responses in the ?SH condition and 57.9% "SH" responses in the ?S condition), $F(1, 62) = 0.29$, $p_{rep} = .44$, $d = 0.12$.

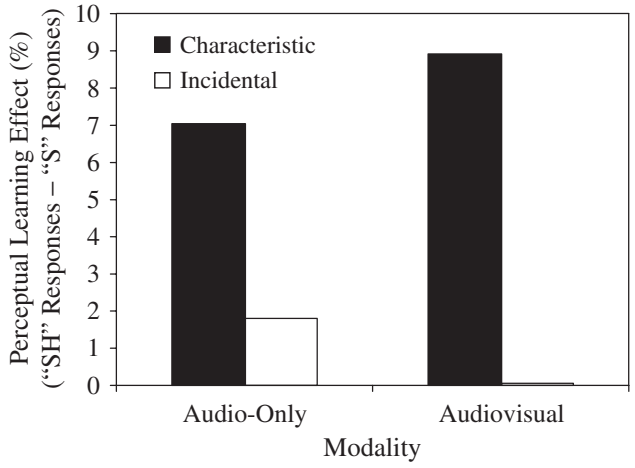

Fig. 1. Perceptual learning (percentage of ''SH'' responses minus percentage of ''S'' responses) as a function of modality of presentation and attribution condition.

The attribution manipulation in the audiovisual condition produced the same pattern of results (see Fig. 3): Participants distinguished characteristic from incidental pronunciations, showing learning only for the former. Those exposed to [∼s∫] when the speaker did not have a pen in her mouth (characteristic) showed robust learning (57.0% "SH" responses in the ?SH condition vs. 48.1% "SH" responses in the ?S condition), $F(1, 68) = 6.29$, $p_{rep} = .94$, $d = 0.6$. In contrast, participants exposed to [∼s∫] when that pronunciation could be attributed to a pen in the mouth (incidental) showed no such learning: Those in the ?SH condition and those in the ?S condition categorized the same number of items as "SH" (60.8% vs. 60.7%, respectively), $F(1, 68) = 0.04$, $p_{rep} = .06$, $d = 0.008$.

DISCUSSION

These findings address the puzzle of how perceptual constancy in speech perception is achieved in the face of pervasive variation: The system integrates available cues about whether a variation is characteristic of the speaker who is producing it or an incidental consequence of some other factor. If the variation seems characteristic, the appropriate phonemic representation is restructured to accommodate it; if the variation seems incidental, no such restructuring occurs. The process of rapidly recognizing and extracting invariance is guided by pragmatic attributions about the variation's source.

Our results provide compelling evidence that perceptual learning provides a mechanism by which the perceptual system can flexibly accommodate idiosyncratic variation without resulting in indiscriminate restructuring. The presence of a nonstandard pronunciation when the system was initially exposed to a speaker caused learning to engage; the same nonstandard pronunciation heard after standard input did not. The pen-in-the-mouth manipulation provided an explicit and temporary excuse for a speaker's nonstandard pronunciation, again blocking learning, and overriding the system's first-impression bias.

That the same acoustic variation can have different perceptual consequences represents a radical departure from the predictions of dominant theories of perceptual learning in speech processing, in which variations are not distinguished on the basis of their source. This finding similarly departs from theories in which listeners must have direct access to the articulatory gestures that produce the signal, as even in those

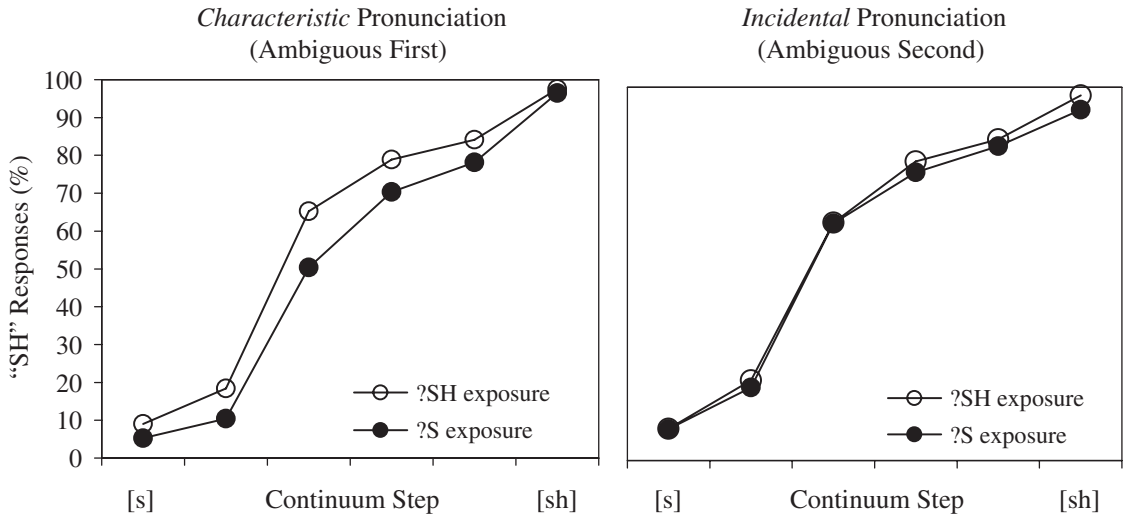

Fig. 2. Percentage of ''SH'' responses for each categorization-test item in the audio-only conditions as a function of perceptual-learning condition (?SH vs. ?S). Results are shown separately for the characteristic-pronunciation and incidental-pronunciation conditions.

theories, the signal itself is the primary source of information for recovering the underlying gesture. Our finding that episodic order biases whether a pronunciation is learned, and that a temporary pragmatic attribution can override this bias to cause learning to proceed more conservatively, has compelling implications for spoken-language comprehension more generally: Linguistic processing is not and cannot be independent of pragmatic factors, such as knowledge about the speaker.

In our formulation, perception and learning are united under a broad conception of "pragmatics": The order effects shown in learning and the context effects that inform perception both stem from pragmatic considerations and are constrained by them. The order effects, in particular, may also be consistent with other (nonpragmatic) mechanisms, such as attentional or maturational constraints (e.g., Elman, 1993); we hope that future work will allow researchers to clarify the influence of multiple constraints in the learning process.

Previous proposals for accommodating pragmatic information do so via one of two mechanisms. In one, the listener must model a speaker and consult that model either continuously (H.H. Clark & Marshall, 1978) or following an initial stage of egocentric processing, during monitoring and repair (e.g., Brown & Dell, 1987; Horton & Keysar, 1996). In the other, no such modeling by the listener is required because perception and

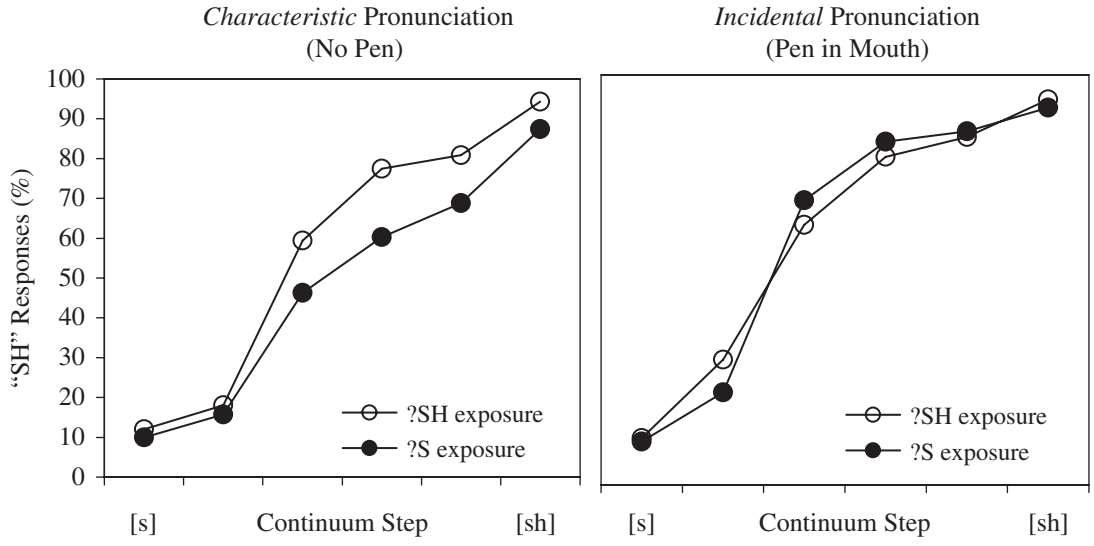

Fig. 3. Percentage of ''SH'' responses for each categorization-test item in the audiovisual conditions as a function of perceptual-learning condition (?SH vs. ?S). Results are shown separately for the characteristic-pronunciation and incidental-pronunciation conditions.

production are assumed to be tightly coupled; speakers' productions are accommodated in a "dumb," resource-free way via "output-input coordination" (Garrod & Anderson, 1987) or imitation (Pickering & Garrod, 2004). But the perceptual-learning mechanism supported by our data requires that pragmatic information be reconceptualized; it can be closely integrated with linguistic information and need not be represented extraneously or used only in a slow, late-occurring inferential stage. When pragmatic information signals what is invariant, it constrains the extent to which perceptual learning occurs.[1] This application of pragmatic information explains why knowledge about a conversational partner sometimes immediately affects perception and sometimes appears not to: Characteristic properties affect representations; incidental ones do not.

Acknowledgments—This material is based on work supported by National Science Foundation Grant 0325188, by National Institutes of Health (NIH) Grants F32 HD052342 and R0151663, and by an NIH-funded postdoctoral training grant at the Center for Research in Language, University of California, San Diego.

REFERENCES

Albert, M.K., & Hoffman, D.D. (2000). The generic-view assumption and illusory contours. *Perception*, *29*, 303–312.

Arnold, J.E., Hudson Kam, C.L., & Tanenhaus, M.K. (2007). If you say thee uh – you're describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 914–930.

Binford, T.O. (1981). Inferring surfaces from images. *Artificial Intelligence*, *17*, 205–244.

Brown, P.M., & Dell, G.S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology*, *19*, 441–472.

Clark, A., & Thornton, C. (1997). Trading spaces: Computation, representation, and the limits of uninformed learning. *Behavioral and Brain Sciences*, *20*, 57–66.

Clark, H.H., & Marshall, C.R. (1978). Reference diaries. In D.L. Waltz (Ed.), *Theoretical issues in natural language processing* (Vol. II, pp. 57–63). New York: Association for Computing Machinery.

Elman, J.L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–99.

Fowler, C.A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, *14*, 3–28.

Freeman, W.T. (1994). The generic viewpoint assumption in a framework for visual perception. *Nature*, *368*, 542–545.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*, 181–218.

Gibson, J.J. (1966). The problem of temporal order in stimulation and perception. *Journal of Psychology*, *62*, 141–149.

Gibson, J.J. (1973). On the concept of 'formless invariants' in visual perception. *Leonardo*, *6*, 43–45.

Horton, W.S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*, 91–117.

Kraljic, T., Brennan, S.E., & Samuel, A.G. (in press). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*.

Kraljic, T., & Samuel, A.G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*, 141–178.

Kraljic, T., & Samuel, A.G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*, 262–268.

Liberman, A., & Mattingly, I. (1985). The motor theory revised. *Cognition*, *21*, 136.

Maye, J., Aslin, R., & Tanenhaus, M. (in press). The weckud wetch of the West: Rapid adaptation to a novel accent. *Cognitive Science*.

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

Pickering, M.J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169–190.

Rock, I. (1983). *The logic of perception*. Cambridge, MA: MIT Press.

(RECEIVED 6/10/07; REVISION ACCEPTED 9/17/07)

[1]This logic applies to language comprehension more generally. Recent work on how adults interpret referring expressions suggests that listeners adjust expectations about a referent on-line when they have a permanent attribution for a speaker's disfluency (the speaker has agnosia), but not when the attribution for disfluency is temporary (the speaker was distracted by noise; Arnold, Hudson Kam, & Tanenhaus, 2007).