Perceptual learning in speech: Stability over time (L)

Frank Eisner[a]
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands and Institute of Cognitive Neuroscience, University College London, United Kingdom

James M. McQueen
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

(Received 23 June 2005; revised 27 January 2006; accepted 31 January 2006)

Perceptual representations of phonemes are flexible and adapt rapidly to accommodate idiosyncratic articulation in the speech of a particular talker. This letter addresses whether such adjustments remain stable over time and under exposure to other talkers. During exposure to a story, listeners learned to interpret an ambiguous sound as [f] or [s]. Perceptual adjustments measured after 12 h were as robust as those measured immediately after learning. Equivalent effects were found when listeners heard speech from other talkers in the 12 h interval, and when they had the opportunity to consolidate learning during sleep. © *2006 Acoustical Society of America.*

[DOI: 10.1121/1.2178721]

PACS number(s): 43.71.Bp, 43.71.Es [ARB] Pages: 1950–1953

I. INTRODUCTION

When we listen to speech, we need to adjust our interpretation of speech cues in response to talker-specific differences in articulation (Ladefoged, 1989; Ladefoged and Broadbent, 1957). The variability in the speech signal that is introduced by talker idiosyncrasies continues to be problematic for automatic speech recognizers, but is usually handled with remarkable ease by the human perceptual system. By comparing comprehension of novel and familiar talkers under difficult listening conditions, Nygaard *et al.* (1994) and Nygaard and Pisoni (1998) have shown that being familiar with a talker's voice can even aid comprehension once an initial adjustment has been made.

There are likely to be various processes engaged in perceptual adjustments made to a talker, driven by different sources of talker variability, and operating at several levels, such as the phonemic, lexical, and prosodic levels. A recent study has shown one specific mechanism, which uses lexical knowledge to resolve ambiguities that arise in the signal at the sublexical level (Norris *et al.*, 2003). Exposure to an ambiguous sound [?], that was midway between [f] and [s], caused a shift of the [f]–[s] category boundary when [?] was placed in contexts that were lexically consistent with its interpretation as either [f] or [s]. Two groups of Dutch listeners heard this ambiguous sound while performing a lexical decision task, either in contexts favoring [f] (e.g., *olij?*, where *olijf* is a word, "olive," but *olijs* is not), or in contexts favoring [s] (e.g., *radij?*, where *radijs* is a word, "radish," but *radijf* is not). Listeners in the first group subsequently categorized more sounds on an [f]–[s] continuum as [f] than listeners in the second group.

The studies by Nygaard *et al.* and Norris *et al.* suggest that the perceptual system has access to previously acquired information about a talker. The present study asks whether this kind of perceptual learning remains stable over a 12-h period. This follows up on recent research using the Norris *et al.* exposure-test paradigm that has shown a solid, and under some conditions even increased, perceptual adjustment effect 25 min after learning (Kraljic and Samuel, 2005). A second question was whether conditions that favor consolidation of learning, such that there is little contact with other talkers, as well as the opportunity for sleep, produce a more robust effect than conditions where participants have normal day-to-day interaction with other talkers, and no sleep. A study in which participants were trained to understand synthetic speech has found that, for this type of learning, there is indeed a performance increase when the testing conditions allow sleep over conditions without sleep (Fenn *et al.*, 2003).

To address these questions, an adapted version of the Norris *et al.* (2003) paradigm was used for inducing a perceptual adjustment. Listeners were first pretested on their categorization of [f]–[s] sounds before having lexically biased exposure to an ambiguous fricative, in the context of listening to a story. They were tested again on [f]–[s] categorization immediately after exposure, and after a 12-h delay, either over the course of one day, or with an intervening night's sleep.

II. METHOD

A. Participants

Eighty-four native Dutch speakers with no self-reported hearing disorders took part in exchange for a cash payment. Twenty-four participated in pretests, and 60 participated in the main experiment.

B. Materials and stimulus construction

Speech recordings were made in a sound-damped booth (Sony ECM-MS957 microphone) in a single session and digitized for further processing (Sony SMB-1 A/D converter; 44.1 kHz sampling rate; 16-bit quantization). A female na-

[a] Electronic mail: f.eisner@ucl.ac.uk

tive Dutch speaker produced 20 tokens each of the syllables [ɛf], [ɛs], and [ɛx] for test stimulus construction, and read out two versions of a story (see below).

1. [ɛf]–[ɛs] continuum

One token each of [f] and [s] was selected from the recorded syllables and excised at zero crossings at the onset of frication (original durations: [s] 246 ms, [f] 234 ms; original intensities: [s] 67.7 dB SPL, [f] 61.3 dB SPL). The fricatives were cut to a duration of 231 ms, and equated in root-mean-square intensity (62.4 dB SPL). With these sounds as endpoints, an 81-step continuum was made by combining their waveforms in graded, equally spaced proportions (effectively manipulating the spectrum; see McQueen, 1991), where step 1 corresponded to a clear [f] and step 81 to a clear [s]. The resulting fricatives were spliced onto a vowel excised from one of the [ɛx] syllables (duration 111 ms; intensity 79.2 dB SPL). The velar vocalic context was used for all spliced sounds in the experiment in order to avoid transitional cues for [f] or [s].

The [ɛf]–[ɛs] continuum was pretested with 24 Dutch listeners in order to find a maximally ambiguous sound for the exposure materials, and to select stimuli for the test phases of the main experiment. First, 12 listeners categorized ten sounds from the ambiguous range of the continuum (between steps 17 and 53; presented ten times each, in pseudorandomized order). Using the same procedure, a further 12 listeners then categorized ten stimuli taken from a narrower ambiguous range as determined by the first group's responses (between steps 30 and 53). From the second group's responses, steps on the continuum corresponding to 90, 70, 50, 30, and 10 percent of [f] responses were identified or determined by interpolation. The resulting steps (25, 34, 43, 52, and 61) were used in the test phases of the main experiment. The most ambiguous sound, step 43 ([?]), was also used to create the materials for the exposure phase.

2. Story

The text of a Dutch translation of a story (de Saint-Exupéry, 2001, Chap. 2) was edited such that it contained an equal number of [f] and [s] sounds and neither of the sounds [v] or [z]. After editing there were 644 words in total, containing 78 [f] sounds and 78 [s] sounds. Two versions of the story were recorded. In one version, every instance of [f] was intentionally mispronounced as the voiceless velar fricative [x] (e.g., *alsof* "as if" → [alsɒx]). In the second version every [s] was pronounced as [x] (e.g., *alsof* → [alxɒf]). The 78 critical velar fricatives in both versions were then excised at zero crossings and replaced by a version of the ambiguous fricative [?]. Since in natural speech the duration of segments is conditioned by various contextual factors, there were three tokens of [?] (all based on step 43). These were made by modifying the amplitude envelope to create two shorter 60-ms and 100-ms sounds (linearly ramped over a 10-ms window at onset and offset), and a long 160-ms sound (ramped over 10 ms at onset and 40 ms at offset). For any given position, the most natural-sounding token out of these three was used. The final two versions of the story were 4.0 min long.

C. Design and procedure

All participants were given a pretest in which they categorized the five [ɛf]–[ɛs] steps, followed by an exposure phase where the task was simply to listen to one of the two story versions. Immediately after exposure, there was a first categorization post-test, and after a delay of 12 h, a second post-test.

For 30 participants, the pretest started at 9 am, and post-test–2 was at 9 pm on the same day ("day group"). For a further 30 subjects, the first session began at 9 pm, while post-test–2 took place at 9 am the following morning ("night group"). In each of those groups, there were 15 listeners who heard the [f]-biased version of the story during exposure (i.e., [?] replacing [f]), and 15 listeners who heard the [s]-biased version.

Pretest, post-test-1, and post-test-2 all consisted of ten randomizations of the same five [ɛf]–[ɛs] steps. Stimuli were presented at an interonset interval of 2600 ms. Listeners were tested in groups of up to four, and instructed to press a button labeled "F" when hearing an [f]-like sound, and a button labeled "S" for an [s]-like sound.

III. RESULTS

For every test phase, listeners' responses were converted to a percentage of [f] categorizations per step. Data from three participants (day group; [s]-biased exposure) were corrupted due to a technical error, and discarded. All listeners in the night groups confirmed having had at least 6 h of sleep between the post-tests.

A. Stability of learning

An initial analysis of variance (ANOVA) with test (pretest, post-test-1, or post-test-2) and step (the five [f]–[s] sounds) as within-subjects factors and lexical bias ([f]- or [s]-biased exposure) as a between-subjects factor revealed a significant interaction of test and lexical bias [$F(2,110)=3.68$, $p=0.028$]. The interaction was examined by conducting ANOVAs with step and lexical bias as factors separately for each test phase. While there was no significant difference between the two exposure groups at pretest [$F(1,55)=0.02$, $p=0.893$; see Fig. 1(a)], their respective categorization functions were significantly different from each other immediately after the exposure phase [$F(1,55)=5.76$, $p=0.020$; see Fig. 1(b)], and after a 12-h delay [$F(1,55)=4.76$, $p=0.033$; see Fig. 1(c)]. For a direct comparison of these perceptual learning effects in post-tests 1 and 2, we conducted an ANOVA with test (post-test-1 or post-test-2), lexical bias, and step as factors. Crucially, the interaction of test and lexical bias was not significant [$F(1,55)=0.13$, $p=0.726$], suggesting that perceptual learning was as robust after 12 h as it was immediately after learning.

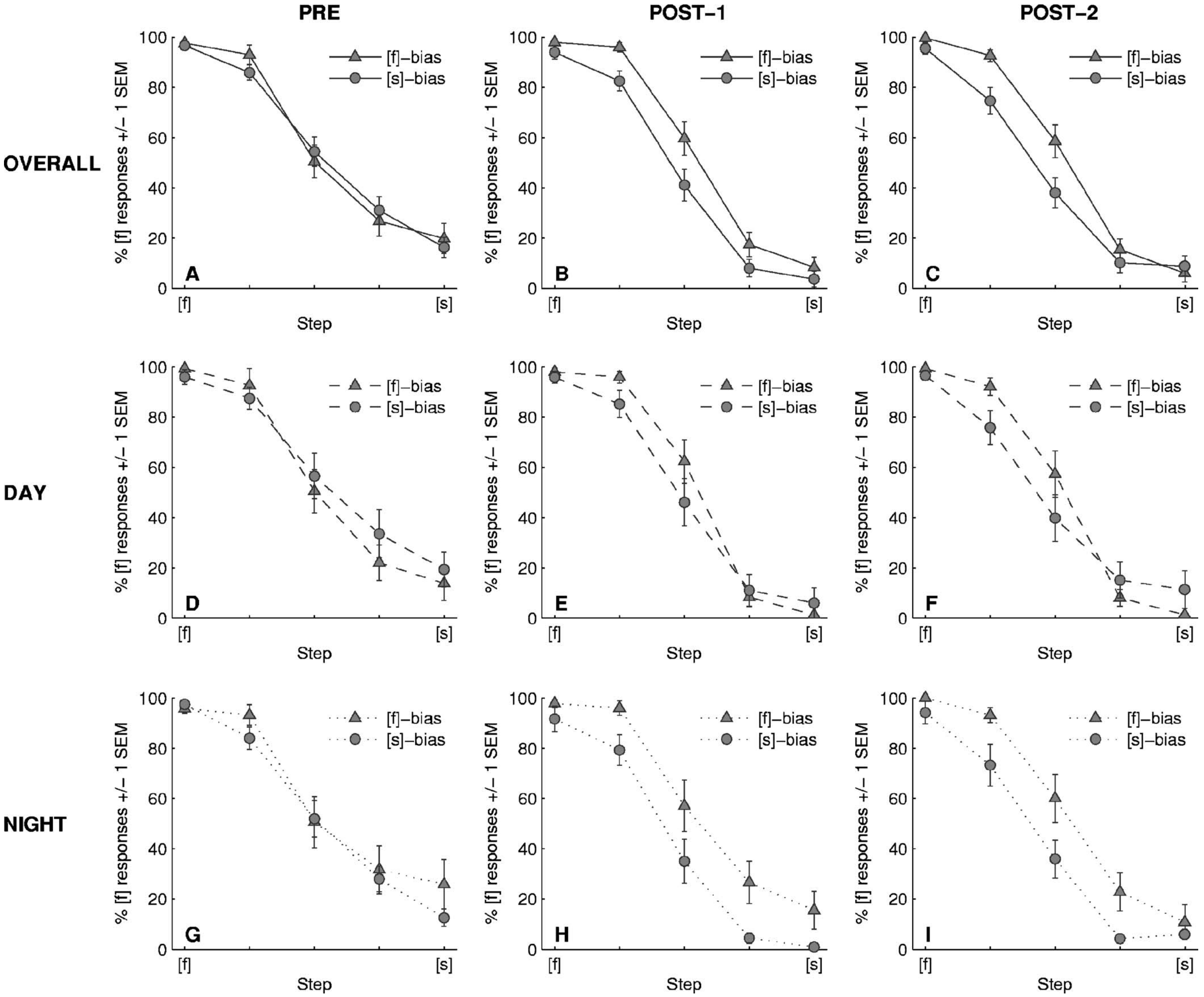

FIG. 1. Percentages of [f] responses to each of the five [f]–[s] steps for the groups with [f]-biased and [s]-biased exposure at pretest, post-test-1, and post-test-2. The top panels (A), (B), (C) show the mean performance collapsed across groups; the middle panels (D), (E), (F) show the day groups only; and the bottom panels (G), (H), (I) show the night groups.

B. Delay at night vs during the day

To test for a specific effect of sleep on post-test-2 performance, a further ANOVA comprised the factors time of exposure (9 am or 9 pm), test (pretest or post-test-2), lexical bias, and step. An effect of sleep on perceptual learning would be reflected in the interaction of time of exposure × test × lexical bias, which was not significant, $F(1,53)=0.15$, $p=0.699$. The test × lexical bias interaction was significant, $F(1,53)=4.20$, $p=0.045$. We also tested for a potential effect of time of exposure on immediate learning by conducting an equivalent ANOVA with the post-test-1 data. Again, the three-way interaction of time of exposure × test × lexical bias was not significant $[F(1,53)=0.26, p=0.615]$, while the overall learning effect was significant [test × lexical bias, $F(1,53)=4.60$, $p=0.037$]. Thus, although the immediate learning effect was numerically stronger in the night group than in the day group (see Fig. 1), the difference between groups was not significant. The effect, once established, remained stable over the 12-h delay, both overnight and during the day, as shown by a lack of interaction of time of exposure × test × lexical bias in an equivalent ANOVA comparing post-tests 1 and 2 $[F(1,53)=0.05, p=0.823]$.

IV. DISCUSSION

The results show an immediate perceptual learning effect after hearing an ambiguous fricative sound [?] in lexically biased contexts for a few minutes. In contrast to previous studies using a lexical decision task on a list of words and nonwords as the exposure phase (Eisner and McQueen, 2005; Kraljic and Samuel, 2005; Norris *et al.*, 2003), this lexically guided learning effect was observed here when exposure was listening to a short story and thus involved no decision task. Listeners who heard the ambiguous sound placed in words that favor its interpretation as an [f] labeled more sounds on an [f]–[s] continuum as [f] than they did before exposure to [?], while listeners who heard the same sound in [s]-biased contexts showed the reverse pattern. The effect remained robust after a 12-h interval: No change in magnitude in either direction was observed (relative to the immediate post-test), both for the groups which had the opportunity for consolidation during sleep and received relatively little speech input from other talkers, and the groups which had no sleep and more contact with other talkers.

Fenn *et al.* (2003) showed that, for learning to understand synthetic speech, there is a decrease in performance during 12 h of waking but subsequent recovery during sleep. The lack of such a pattern in the present data suggests that the type of perceptual learning examined here is less susceptible to decay. In contrast to learning about synthetic speech, a perceptual adjustment to a talker idiosyncrasy is a very fast-occurring process in which listeners already are highly skilled, and of which they are therefore usually unaware. The perceptual system in this case is not learning a novel skill as such, but applying a subtle adjustment in the processing of a particular phoneme contrast. For this kind of learning to be helpful to the listener in benefiting subsequent recognition of the exposure talker's speech (Norris *et al.*, 2003), it ought to occur rapidly and remain stable, regardless of whether the listener is awake or asleep. Although learning to understand synthetic speech better presumably taps into existing prelexical adjustment routines, it is likely to also involve learning at other processing levels (e.g., the unusual prosody of the synthetic "talker"), all of which may be subject to unlearning during waking. This type of learning also takes time and effort (Greenspan *et al.*, 1988), and often requires explicit feedback during training. It is therefore quite possible that a more drastic distortion of the natural speech signal than the manipulation in the present experiment (e.g., affecting more than one phoneme contrast, or additional levels of processing) will also be more liable to the process of unlearning and recovery that Fenn *et al.* have demonstrated for synthetic speech.

The picture that is emerging for lexically driven perceptual adjustments in response to talker idiosyncrasies is that these remain very stable. Using a similar paradigm as the present study, Kraljic and Samuel (2005) have already shown that learning effects are reliable after a 25-min interval, unless listeners are exposed to unambiguous tokens of the critical sound that come from the voice of the exposure talker. Together with these results, the evidence at present suggests that, once the perceptual system has adjusted to a given talker, it does not return to its original state through either the effects of speech input from other talkers or the mere passage of time.

ACKNOWLEDGMENTS

We would like to thank Femke Verhees for help with constructing the stimulus materials and Dennis Pasveer for testing some of the participants. We also thank Arty Samuel and an anonymous reviewer for comments on a previous version of this manuscript.

de Saint-Exupéry, A. (**2001**). *De Kleine Prins* (Ad Donker, Rotterdam [original work published **1943**]).

Eisner, F., and McQueen, J. M. (**2005**). "The specificity of perceptual learning in speech processing," Percept. Psychophys. **67**(2), 224–238.

Fenn, K. M., Nusbaum, H. C., and Margoliash, D. (**2003**). "Consolidation during sleep of perceptual learning of spoken language," Nature (London) **425**, 614–616.

Greenspan, S. L., Nusbaum, H. C., and Pisoni, D. B. (**1988**). "Perceptual learning of synthetic speech produced by rule," J. Exp. Psychol. Learn. Mem. Cogn. **14**(3), 421–433.

Kraljic, T., and Samuel, A. G. (**2005**). "Perceptual learning for speech: Is there a return to normal?" Cogn. Psychol. **51**, 141–178.

Ladefoged, P. (**1989**). "A note on 'Information conveyed by vowels'," J. Acoust. Soc. Am. **85**, 2223–2224.

Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 98–104.

McQueen, J. M. (**1991**). "The influence of the lexicon on phonetic categorization: Stimulus quality in word-final ambiguity," J. Exp. Psychol. Hum. Percept. Perform. **17**, 433–443.

Norris, D., McQueen, J. M., and Cutler, A. (**2003**). "Perceptual learning in speech," Cogn. Psychol. **47**, 204–238.

Nygaard, L. C., and Pisoni, D. B. (**1998**). "Talker-specific learning in speech perception," Percept. Psychophys. **60**, 355–376.

Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (**1994**). "Speech perception as a talker-contingent process," Psychol. Sci. **5**(1), 42–46.