Journal of Experimental Psychology: General
2005, Vol. 134, No. 2, 222–241

0096-3445/05/$12.00 DOI: 10.1037/0096-3445.134.2.222

Lexical Information Drives Perceptual Learning of Distorted Speech: Evidence From the Comprehension of Noise-Vocoded Sentences

Matthew H. Davis
MRC Cognition and Brain Sciences Unit

Ingrid S. Johnsrude
MRC Cognition and Brain Sciences Unit
and Queen's University

Alexis Hervais-Adelman and Karen Taylor
MRC Cognition and Brain Sciences Unit

Carolyn McGettigan
University College London

Speech comprehension is resistant to acoustic distortion in the input, reflecting listeners' ability to adjust perceptual processes to match the speech input. For noise-vocoded sentences, a manipulation that removes spectral detail from speech, listeners' reporting improved from near 0% to 70% correct over 30 sentences (Experiment 1). Learning was enhanced if listeners heard distorted sentences while they knew the identity of the undistorted target (Experiments 2 and 3). Learning was absent when listeners were trained with nonword sentences (Experiments 4 and 5), although the meaning of the training sentences did not affect learning (Experiment 5). Perceptual learning of noise-vocoded speech depends on higher level information, consistent with top-down, lexically driven learning. Similar processes may facilitate comprehension of speech in an unfamiliar accent or following cochlear implantation.

Humans are able to understand speech in a variety of situations that dramatically affect the sounds that reach their ears. They can understand talkers with quite different (foreign or regional) accents, who speak at different speeds, in rooms that introduce reverberation, or when the speech is conveyed over low-fidelity devices such as the telephone. The robustness of speech comprehension to many forms of variation and distortion is currently unmatched by computer speech recognition systems and may therefore reflect a unique specialization of the human perceptual system.

Experimental work has demonstrated that speech perception remains robust even when challenged with extreme forms of artificial distortion. For example, speech remains understandable when formants are resynthesized as sinusoids (Remez, Rubin, Berns, Pardo, & Lang, 1994; Remez, Rubin, Pisoni, & Carrell, 1981), a manipulation that removes most of the natural qualities of the human voice from a speech signal. Other manipulations have shown that dramatic alterations to both the temporal (Mehler et al., 1993; Saberi & Perrott, 1999) and spectral (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995; Smith, Delgutte, & Oxenham, 2002) properties of speech do not substantially impair the intelligibility of spoken language when background noise is absent.

It is unlikely that there is any single set of acoustic properties or cues that are preserved in all of these different forms of distorted yet still-intelligible speech (cf. Bailey & Summerfield, 1980). Therefore, robustness in speech perception reflects the multiple acoustic means by which stable elements of speech (such as phonetic features or syllables) are coded in clear speech: This redundancy permits comprehension when any single cue is lost. Robustness in speech comprehension may also derive from the operation of compensatory mechanisms that are recruited when speech becomes difficult to understand: processes of adaptation and perceptual learning are two such mechanisms.

The human language system can dynamically adapt to variation in the acoustic realization of speech, tuning the perceptual system so as to optimally process the current speech input. These adaptation processes can take place very rapidly. For example, adaptation to natural changes in speech rate (e.g., J. L. Miller, 1981; J. L. Miller & Lieberman, 1979; Summerfield, 1981) or to changes in the spectral characteristics of the communication channel (Ladefoged & Broadbent, 1957; Summerfield, Haggard, Foster, & Gray, 1984; Watkins, 1991) occur in less than a second. Such adjustments occur relatively automatically and largely without listeners being aware of the perceptual consequences of their operation. However, not all changes in the perception of spoken input can be characterized as rapid and effortless adaptation.

More extreme or unnatural forms of distorted speech require longer periods of exposure for listeners to achieve full comprehension. For example, adaptation to artificial changes in speech

Matthew H. Davis, Alexis Hervais-Adelman, and Karen Taylor, MRC Cognition and Brain Sciences Unit, Cambridge, United Kingdom; Carolyn McGettigan, Department of Human Communication Science, University College London, London, United Kingdom; Ingrid S. Johnsrude, MRC Cognition and Brain Sciences Unit, Cambridge, United Kingdom, and Department of Psychology, Queen's University, Kingston, Ontario, Canada.

Additional materials are on the Web at http://dx.doi.org/10.1037/0096–3445.134.2.222.supp. Examples of noise-vocoded sentences can also be found at http://www.mrc-cbu.cam.ac.uk/~matt.davis/vocode/.

Correspondence concerning this article should be addressed to Matthew H. Davis, MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 2EF, United Kingdom. E-mail: matt.davis@mrc-cbu.cam.ac.uk

rate that are outside the normal range (as in time-compressed speech) may require several minutes of compressed input (Mehler et al., 1993; Pallier, Sebastian-Galles, Dupoux, Christophe, & Mehler, 1998). We consider that a form of long-term learning is required for listeners to change their perception in such cases, as evidenced by the fact that trained listeners outperform naive listeners when tested 1 year after their original exposure to time-compressed speech (Altmann & Young, 1993). We will refer to these processes as perceptual learning because they reflect long-term changes in a listener's ability to extract information from speech input. This conforms to Goldstone's (1998, p. 586) definition of perceptual learning as involving "relatively long-lasting changes to an organism's perceptual system that improve its ability to respond to its environment and are caused by this environment." Multiple mechanisms can effect perceptual learning: what they have in common is that they result in improved identification and discrimination for the perceptual dimension involved (Fahle & Poggio, 2002; Goldstone, 1998).

Experimental investigations suggest that perceptual learning processes are invoked in many situations in which listeners are faced with speech sounds that are outside the range that they have previously experienced. For instance, when presented with heavily accented speech, listeners must adjust to a set of unfamiliar phonemic and prosodic properties. Experiments have demonstrated that effective perception of speech in an unfamiliar accent can require several minutes or more of exposure to allow full comprehension (Bent & Bradlow, 2003; Clarke & Garrett, 2004; Weill, 2001).

Experimental investigations of artificial distortions that simulate the real-world variability encountered in accented speech have been used to explore the nature of the learning process in more detail. In a recent series of experiments, Norris and colleagues (Norris, McQueen, & Cutler, 2003) exposed listeners to artificially modified speech in which an ambiguous fricative replaced all occurrences of either /f/ or /s/, simulating changes in phoneme boundaries typical of accented speech. Following exposure to 20 words that contained this ambiguous fricative, listeners subsequently showed a marked difference in their perception of /s/ and /f/, depending on which phoneme had been altered during training. Perceptual changes were only observed for listeners exposed to the ambiguous fricative in the context of real words but not in nonwords. In this case, lexical knowledge seems to play a crucial role in learning the correct interpretation of an ambiguous fricative. These findings offer an intriguing glimpse of a role for higher level information in supporting lower level perceptual processes. As in other perceptual domains (e.g., vision; Pylyshyn, 1999), this is a topic of considerable debate in cognitive accounts of speech perception (see Norris, McQueen, & Cutler, 2000; Samuel, 1997, 2001).

Models of speech perception typically postulate a number of processing stages that mediate between acoustic analyses of the speech signal and higher level representations of the meaning of an utterance (Gaskell & Marslen-Wilson, 1997; Luce & Pisoni, 1998; McClelland & Elman, 1986; Norris, 1994). Although the units proposed at intervening levels vary between accounts (e.g., phonetic features, phonemes or syllables at a sublexical level, morphemes, words or meaning at a lexical level), there is agreement that recognition proceeds in hierarchically organized stages, with greater abstraction from the surface details of speech at higher levels. Although the extent of this abstraction from the surface properties of speech (particularly at the lexical level) remains a topic of debate (Goldinger, 1998), the majority of models intended to account for word recognition phenomena are abstractionist and suggest that utterance-specific or voice-specific details are not retained at a lexical level (Gaskell & Marslen-Wilson, 1997; Luce & Pisoni, 1998; McClelland & Elman, 1986; Norris, 1994). In order to account for the apparent ease with which acoustically variable productions of spoken words can be recognized, abstractionist accounts typically propose mechanisms of perceptual normalization to compensate for variations in the acoustic realization of speech (Pisoni, 1997). Experimental evidence suggests that these normalization mechanisms do not only involve low-level processes but also potentially affect higher level processes (Nusbaum & Magnuson, 1997). At present, however, there are few well-specified accounts of the mechanisms involved in tuning the perceptual system to incoming speech and few experimental techniques for establishing whether this normalization is dependent on or independent of higher level, lexical information.

In this article, we present a novel experimental approach to these issues and investigate the influence of low-level (acoustic–phonetic) and high-level (lexical) factors on learning to understand artificially distorted speech. We use noise-vocoded speech, which is an acoustic distortion that preserves temporal information while removing the temporal fine structure and spectral detail of speech (Shannon et al., 1995). Although initially unintelligible, noise-vocoded sentences can be readily understood following a period of training. This artificial distortion therefore provides a model system in which to conduct experimental investigations of the processes involved in adjusting to novel-sounding speech.

Noise-Vocoded Speech

Noise-vocoded speech is created by dividing the speech signal into frequency bands (analogous to the individual electrodes in a cochlear implant) and then applying the amplitude envelope in each frequency range to band-limited noise (producing a dramatic loss of spectral detail). These processing steps are depicted on spectrograms of an example sentence in Figure 1. The result of this procedure is a stimulus (noise-vocoded speech) that sounds like a harsh, noisy whisper. Sample noise-vocoded sentences can be found in supplemental materials on the Web at http://dx.doi.org/10.1037/0096–3445.134.2.222.supp, and further examples with different numbers of frequency bands can be found at http://www.mrc-cbu.cam.ac.uk/~matt.davis/vocode/.

The eventual intelligibility of noise-vocoded sentences has been shown to depend on a number of properties of the vocoded stimulus. For instance, increasing the number of frequency bands in the vocoder improves measured intelligibility (see Davis & Johnsrude, 2003; Loizou, Dorman, & Tu, 1999; Shannon et al., 1995). Speech synthesized with more than 10 bands is readily intelligible, even to entirely naive listeners. Speech vocoded with just four bands can also be highly intelligible (Shannon et al., 1995), although it might require several hours of training for listeners to achieve as high a level of performance as can be rapidly obtained with more bands. The effect that training has on report scores is widely acknowledged; participants in studies using

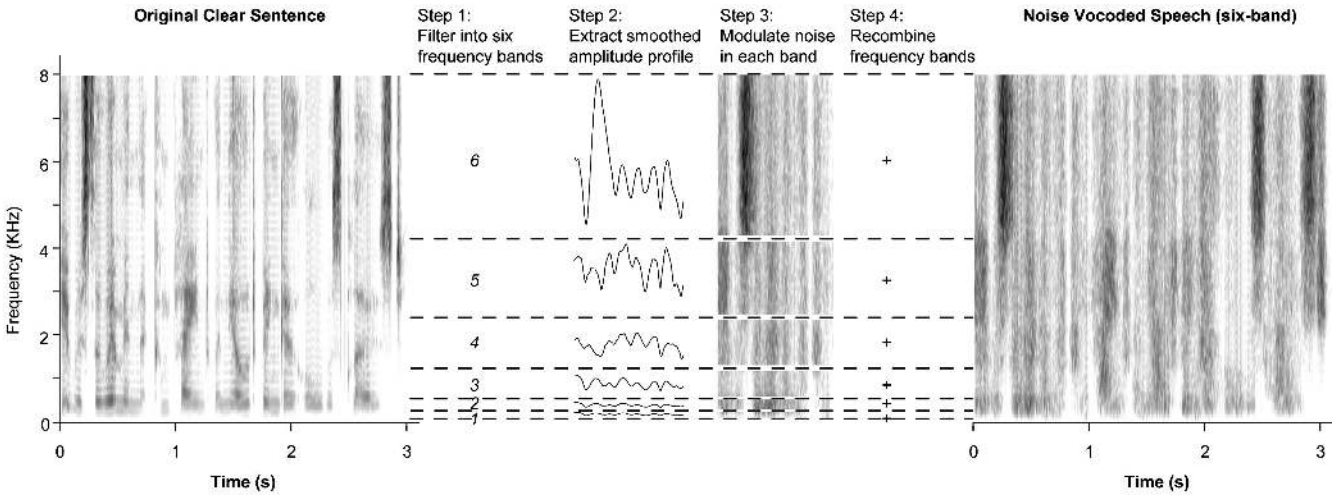

Figure 1. Processing steps involved in transforming clear speech (left spectrogram) into noise-vocoded speech. Sentences are filtered into six nonoverlapping frequency ranges (Step 1), the amplitude envelope in each band is extracted and smoothed (Step 2), and wide-band noise in each frequency range is modulated using this amplitude envelope (Step 3) and combined to produce a noise-vocoded sentence (Step 4; right spectrogram).

vocoded speech typically receive a period of training before performance is assessed (e.g., Baskent & Shannon, 2003; Scott, Blank, Rosen, & Wise, 2000; Shannon et al., 1995). Recent work has tracked performance throughout the learning process (Fu & Galvin, 2003; Rosen, Faulkner, & Wilkinson, 1999). However, to our knowledge, there has been no systematic attempt to determine the characteristics of the learning processes that allow this form of distorted speech to be understood.

One important application for research using noise-vocoded speech is to simulate speech transduced by a cochlear implant (Faulkner, Rosen, & Smith, 2000; Loizou et al., 1999; Shannon et al., 1995). An increased understanding of the processes by which normally hearing listeners learn to understand vocoded speech cannot only inform cognitive accounts of perceptual normalization in speech perception but may also have important implications for users of cochlear implants (see Rosen et al., 1999, for similar arguments).

Learning in Cochlear Implant Users

A cochlear implant is an array of electrodes, implanted into the inner ear, that directly stimulates the auditory nerve in order to restore auditory experience to individuals with a peripheral hearing loss (see Moore & Carlyon, in press; Rauschecker & Shannon, 2002). Following implantation, a prolonged period of rehabilitation (several weeks or months) is typically required for cochlear implant users to achieve optimal comprehension of spoken language (Clark, 2002; Dorman, Hannley, Dankowski, Smith, & McCandless, 1989; Tyler & Summerfield, 1996), and final optimal performance varies widely (Skinner, 2003). Furthermore, whenever changes are made to the way in which sound is conveyed to the electrode array, comprehension suffers and subsequently recovers, reflecting a further period of perceptual adjustment (Fu, Shannon, & Galvin, 2002; Pelizzone, Cosendai, & Tinembart, 1999). Despite this clinical relevance, however, relatively little is known about the auditory learning processes that allow cochlear implant users to understand speech transduced by their implant.

Speech transduced by a cochlear implant retains amplitude envelope information but with reduced spectral detail and altered temporal fine structure. It is therefore well simulated by noise vocoding (Shannon et al., 1995). Research using simulations of cochlear implant-processed speech has thus far focused on the intelligibility of different forms of distorted speech, for instance, simulating the effect of implants with different numbers and placements of electrodes in normal hearing listeners (Dorman, Loizou, & Rainey, 1997; Faulkner et al., 2000; Shannon, Zeng, & Wygonski, 1998). This work provides valuable data on the information that must be conveyed in order for speech to be correctly perceived. However, just as the level of comprehension that is ultimately possible from simulations of cochlear implant-processed speech depends on listeners' ability to adjust to this novel sounding speech (Fu & Galvin, 2003; Rosen et al., 1999), the overall success of cochlear implantation may similarly depend on the ability of implanted patients to learn from their experience of speech as conveyed by their implant. By studying the process by which normal listeners learn to understand noise-vocoded simulations of cochlear implant speech, we may gain insights into how to optimize learning by cochlear implant users, thereby assisting in their rehabilitation.

The Current Study

In a series of five experiments, we manipulated the conditions in which different groups of listeners are exposed to noise-vocoded sentences and measured the consequences of these different training procedures for the eventual intelligibility of distorted sentences. Our aim was to characterize the processes that are respon-

sible for changes in the intelligibility of noise vocoded speech. In particular, we asked whether perceptual changes result from adaptation at early stages of processing—that is, whether learning depends only on listeners being exposed to the sound of noise-vocoded speech—or whether learning also depends on higher level systems involved in understanding and remembering spoken sentences.

In Experiment 1, we examined how comprehension, measured as the proportion of words correctly reported from each sentence, improves in naive listeners over the course of exposure to 30 six-channel noise-vocoded sentences. The results provide a background to subsequent experiments in which we examined the effect of different training conditions on sentence-report scores. Experiments 2 and 3 investigated whether the provision of information on sentence content (in spoken or written form) facilitates learning. Providing the content of each vocoded sentence before it is presented introduces a substantial change in a listener's perceptual experience, because previously unintelligible, distorted sentences become highly intelligible (cf. Giraud et al., 2004; Jacoby, Allan, Collins, & Larwill, 1988; Remez et al., 1981). In these experiments, we assessed whether manipulations that produce this change in perceptual experience also enhance learning and whether the information that produces this enhanced learning is auditory (and hence only present in clear speech) or can also be provided by written presentation (Experiment 3). Experiments 4 and 5 explored the role of higher level (lexical, semantic, and syntactic) information on perceptual learning of speech. Comprehension of noise-vocoded English sentences was tested after an initial period of exposure to noise-vocoded sentences in which lexical, syntactic, or sentential semantic content was systematically manipulated. Comparisons with naive listeners and with listeners trained with vocoded (real) English sentences allow us to determine which aspects of sentence content are crucial for learning, thereby testing whether learning is dependent on the provision of lexical or other higher level information.

Experiment 1: Tracking Changes in the Intelligibility of Noise-Vocoded Speech

Experiment 1 demonstrated the basic effect that lies at the heart of these studies: that report scores for noise-vocoded sentences improve over time. Although it is well-known that the intelligibility of noise-vocoded speech (as for other distortions) changes with practice (see, e.g., Giraud et al., 2004; Rosen et al., 1999; Scott et al., 2000), to our knowledge, the rate and extent of these changes in intelligibility have not been quantified. This experiment also sets out methods that will be used in subsequent experiments.

Method

Participants. Twelve participants from the MRC Cognition and Brain Sciences Unit (Cambridge, United Kingdom) participant panel were tested. All were aged between 18 and 25 years, and most were students at Cambridge University. All were native speakers of British English with no history of hearing or language impairment.

Design and materials. Three groups of 10 vocoded sentences (designated A, B, and C) were presented for report in this experiment. Sentences within each group were presented in a fixed order to each participant. However, the order of the blocks was counterbalanced such that half of the participants heard Group A followed by Group B, whereas the other 6 participants heard Group B followed by Group A. This allows us to compare performance on the first and second blocks of sentences without any confound produced by differences in the difficulty of the various sentences. All participants heard Group C last of all. Report scores for this sentence group provide a stable measure of posttraining performance without sentence-specific effects.

Each group contained 10 simple, declarative sentences with a range of lengths (6 to 13 words, $M = 8.7$ words/sentence) and acoustic durations (1.1–3.0 s, $M = 2.0$ s; see sentence list in Appendix A, which is on the Web at http://dx.doi.org/10.1037/0096–3445.134.2.222.supp). Sentences in each group were equated for length and duration and were matched for naturalness and imageability (rated on 7-point Likert scales by two groups of 18 participants; see Rodd, Davis, & Johnsrude, in press). Sentences with similar properties were used for a four-item memory test and for an example (15-band) vocoded sentence that preceded the main experiment.

The sentences were recorded by a female native speaker of Southern British English in a soundproofed booth. The original recordings were made on digital audiotape, digitally transferred to a Windows PC using a Digital Audio Labs (Chanhassen, MN) Card D sound card and then downsampled to a 22-kHz sampling rate using CoolEdit software (Adobe Systems, San Jose, CA). The recorded sentences were noise vocoded with Praat software (Boersma & Weenink, 2000) using a modified version of a script written by Darwin implementing the processing steps depicted in Figure 1. The sentences were first filtered into six logarithmically spaced frequency bands between 50 and 8000 Hz. Contiguous band-pass filters were constructed in the frequency domain: passbands were 3 dB down at 50, 229, 558, 1161, 2265, 4290, and 8000 Hz with a roll-off of 22 dB/octave (cut-off frequencies chosen to simulate equal distances along the basilar membrane; Greenwood, 1990). The amplitude envelope from each frequency band was extracted using the standard Praat algorithm (squaring intensity values and convolving with a 64-ms Kaiser-20 window, removing pitch-synchronous oscillations above 50 Hz). The resulting envelope was then applied to band-pass filtered noise in the same frequency ranges. Finally, the resulting bands of modulated noise were recombined to produce the distorted sentence. The distorted sentences (along with recorded instructions, silent intervals, and warning tones) were transferred onto an audio CD for presentation to participants in the experiment.

Procedure. Participants were tested in groups of 2 or 3 in a quiet room. All stimuli were played from a Sony portable CD player, through a QED (Veda Products, Bishop's Stortford, Hertfordshire, United Kingdom) headphone amplifier and a splitter box connected to three sets of Sennheiser (Wedemark, Germany) HD25SP headphones. Participants were instructed to listen carefully and write down as many words as they could from each vocoded sentence into an answer book. Before the start of the test, participants had to provide written report for four sentences presented as clear speech. This allowed us to confirm that short-term memory (STM) capacity does not limit participants' performance on the vocoded test sentences. Following the memory test, participants listened to a sample sentence vocoded using 15 bands (which is easily intelligible even to naive listeners) as an example of the form of distortion to be used in the test. They were then presented with 30 vocoded sentences for report. Immediately after each sentence, participants were provided with 25 s in which to write down as much as they could of the sentence that they had heard. A melodic sound instructed participants to stop writing and prepare for the next vocoded sentence. The entire testing session (memory test, example sentence, and vocoded test sentences) lasted approximately 20 min.

Results

Participants' written reports of the clear sentences in the memory test and the vocoded sentences in the main experiment were scored for the percentage of words in each sentence that were

reported correctly. Words were scored as correct only if there was a perfect match between the written form and the word produced in the sentence (morphological variants were scored as incorrect, but homonyms, even if semantically anomalous, were scored as correct). Words were not scored as correct if they were reported in the wrong order, but words reported in the correct order were scored as correct even if intervening words were absent or incorrectly reported. All participants reported all of the words in the memory-test sentences correctly, indicating that STM capacity was not a limiting factor for their report scores. However, report scores for the vocoded sentences were substantially lower—reflecting the increased difficulty of identifying words in distorted speech.

Improvement in report scores over the course of the 30 sentences is evident in Figure 2. Mean report score (across participants) correlates significantly with sentence number in each test order, ABC: $r(30) = .501$, $p < .01$; BAC: $r(30) = .395$, $p < .05$, as illustrated by the trend lines in Figure 2. However, it is also apparent from Figure 2 that report scores vary substantially from item to item. Put simply, some sentences are more difficult to report from vocoded speech than others.

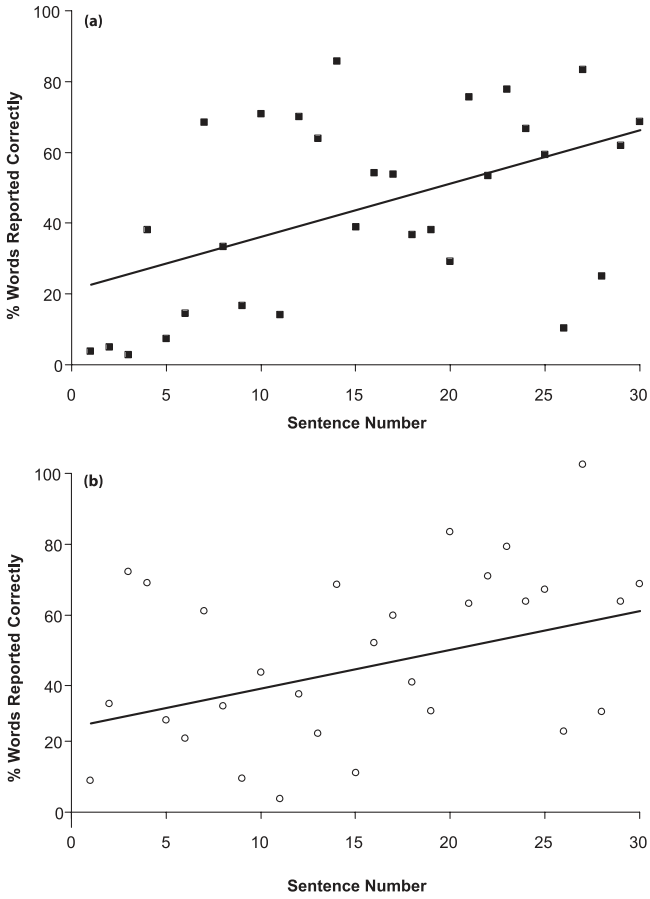

Figure 2. Report scores from 30 noise-vocoded sentences in Experiment 1 averaged over the 6 participants tested on Sentence Group A followed by B and C (a) or Sentence Group B followed by A and C (b). The straight line on each graph shows the best fitting linear relationship between sentence number and report score.

To confirm that changes in report score over the course of the experiment do not result from these differences in sentence difficulty, we conducted analyses of variance (ANOVAs) on report scores from the first 20 sentences grouped into two blocks of 10 sentences and averaged over participants and items. Because two groups of participants were tested on the same sentence blocks in different orders, analysis by items allows us to compare performance on the same sentence when it occurs in either the first or second block of the experiment. An additional dummy variable in the items analysis codes for which group of participants was tested first on each group of sentences, although main effects and interactions involving this dummy variable will not be reported (cf. Pollatsek & Well, 1995). In the analysis by participants, we averaged report scores over the groups of 10 sentences and then compared performance for the first and second block of sentences. Once again, an additional dummy variable codes for the order in which the two sentence groups were presented, although effects of this variable will not be reported (Pollatsek & Well, 1995).

ANOVAs by participants (F_1) and items (F_2) confirmed that there were reliable differences in performance between the first and second group of sentences, $F_1(1, 10) = 12.78$, $MS = 820$, $\eta^2 = .561$, $p < .01$; $F_2(1, 18) = 10.82$, $MS = 1366$, $\eta^2 = .375$, $p < .01$. Report scores improved significantly between the first 10 sentences (31.6%) and the second 10 sentences (43.4%) in the experiment, even when differences in difficulty among items are cancelled out.

Discussion

The results of this experiment demonstrate that report scores on noise-vocoded sentences improve rapidly. Participants are able to report less than 10% of words in the first vocoded sentence that they hear. However, only a few minutes later, participants can report many more of the words in each sentence correctly. Over the course of a 30-sentence exposure, the way in which listeners perceive noise-vocoded speech changes, producing an increase in sentence report scores. The task, writing down the words that were understood on each trial, was an easy and natural task for participants and was performed with 100% accuracy for nondistorted sentences. Therefore, we can be sure that improvement in report scores over time results from changes in participants' perceptual abilities and not from practice with the reporting task. Nevertheless, observing an increase in the number of words reported from vocoded utterances does not inform about what cognitive processes are changing to permit improved comprehension of vocoded speech.

The remaining experiments examined more closely the process that allows report scores for noise-vocoded speech to improve over time. The methods used involve manipulating the conditions under which previously naive listeners are exposed to vocoded speech and measuring the report scores that result from this exposure. Experimental conditions that succeed or fail to produce robust improvements in sentence report scores will provide information about the cognitive processes involved in learning to understand vocoded speech. The first experiment tested whether the improvement arises from a bottom-up process that is dependent only on

exposure to vocoded speech or is assisted by knowledge of the clear sentences that are being presented in vocoded form.

Experiment 2: Effects of Feedback on Learning Vocoded Speech

One striking property that vocoded speech shares with other forms of artificially distorted speech (and speech in noise; Jacoby et al., 1988) such as sine-wave speech (Remez et al., 1981), as well as heavily accented speech, is that even while it is objectively unintelligible (such that few words can be spontaneously reported) the perceived intelligibility of the speech signal can be dramatically altered by information on the content of the sentence. This can be most clearly demonstrated by listening to a vocoded sentence immediately before and immediately after hearing the same sentence as clear speech. The perceptual experience that results from the second presentation is that the vocoded sentence seems to be dramatically more intelligible than on initial hearing. It is as if the content of the sentence pops out from what would otherwise be heavy distortion. Readers are encouraged to experience this for themselves, by listening to demonstration sentences available on the Web at http://dx.doi.org/10.1037/0096–3445.134.2.222.supp. Further examples of noise-vocoded speech can be found at http://www.mrc-cbu.cam.ac.uk/~matt.davis/vocode/.

Experimental investigations of sentences presented in background noise provide a simple demonstration of the effect of stimulus repetition on the perceived clarity of the speech signal. Subjective ratings of the strength of background noise are substantially reduced for sentences (and voices) with which listeners are familiar (Goldinger, Kleider, & Shelley, 1999; Jacoby et al., 1988). Similar phenomena may also be observed in other sensory modalities. For instance, Ahissar and Hochstein (2004) described a "Eureka" experience when viewers perceive degraded visual images before and after presentation of a clear version of the same image. This is quantified in the Gollin (1960) Incomplete Figures Test (Warrington & Weiskrantz, 1968). Familiar drawings can be identified in a more impoverished form than unfamiliar drawings, demonstrating that stimulus repetition reduces the amount of sensory information required for identification. The same is true for noise-vocoded speech; once the clear version of a sentence has been heard and identified, participants can recognize the words in a noise-vocoded sentence much more easily (Giraud et al., 2004). This change in perception we shall refer to as *pop-out*.

This phenomenon of pop-out for vocoded sentences illustrates one way in which a listener's perceptual experience of distorted sentences is a combination of information in the speech signal and higher level influences such as knowledge of the content of that sentence (cf. Remez et al., 1994). Under extreme forms of distortion, this top-down support can produce dramatic changes to the percept evoked by a vocoded sentence. Training also produces a change in the way in which vocoded speech is perceived. It is possible that additional presentations that produce pop-out may influence the learning process. That is, presenting vocoded speech when listeners already know the identity of the sentence not only affects perception of the current sentence but may also assist in the comprehension of subsequent, different sentences.

In testing whether information on the content of distorted sentences affects the learning process, we must ensure that other aspects of participants' experience of vocoded speech are matched. To this end, we compared two groups of participants who reported the same 30 noise-vocoded sentences used in Experiment 1. In this study, however, after reporting from each vocoded sentence, participants heard two repetitions of the test sentence. The first group of participants heard the same sentence presented as clear speech and then as vocoded speech. This distorted–clear–distorted (DCD) condition provided listeners with the experience of hearing each vocoded sentence after the identity of the vocoded sentence is known (from the clear presentation), producing a dramatic increase in the perceived clarity of vocoded sentences on second presentation (pop-out). This DCD condition was compared with a distorted–distorted–clear (DDC) condition in which, after report, participants heard the same number of repetitions of each sentence but without the same experience of pop-out. Participants in the DDC condition only ever had imperfect knowledge of the sentence (based on the words that they could report from the first presentation) before hearing the repetition of the vocoded speech. Even if report scores are high, knowledge of the content of the second vocoded presentation would not be as confident as in the DCD condition.

Method

Participants. Forty-four participants, aged between 18 and 25 years and students at Cambridge University, were tested. All were native speakers of British English with no history of hearing or language impairment. None had taken part in Experiment 1.

Design and materials. The same three groups of 10 sentences (A, B, and C) were used as in Experiment 1, presented in either ABC or BAC order, to counterbalance and thereby control for effects of sentence difficulty. Participants were randomly assigned to one of the two test conditions (DCD and DDC) and one of the two sentence orders.

Procedure. Participants were tested using the same equipment and procedure as in Experiment 1. After writing down what they could understand of each vocoded sentence (at the end of the 25-s delay), participants were cued by a melodic sound to turn over the page in their answer book. Participants were then presented with two repetitions of each sentence—depending on condition, either vocoded and then clear speech (DDC) or vice versa (DCD). Finally, a short tone cued them to expect the next vocoded sentence for report. Participants were instructed to listen carefully to the repetitions of each sentence but were supervised to ensure that they reported only the first presentation of the vocoded sentence and did not alter their response on hearing subsequent repetitions.

Results

Results were scored in the same way as for Experiment 1. Report scores averaged over each condition for the three groups of sentences are shown in Figure 3. Initial analyses concentrated on comparisons of the first two blocks of sentences, in which sentence effects can be cancelled out (as in the analysis of Experiment 1). Scores averaged over participants and items were entered into two-way ANOVAs, with a within-group factor of block (comparing the first and second group of 10 sentences). Condition (DCD vs. DDC) was entered as a between-groups factor by participants and a within-group factor by items. Additional dummy factors (sentence group or sentence order) were added to each analysis as before, although effects of these variables will not be reported.

As can be seen in Figure 3, participants reported more words correctly from the second group of 10 sentences than from the first

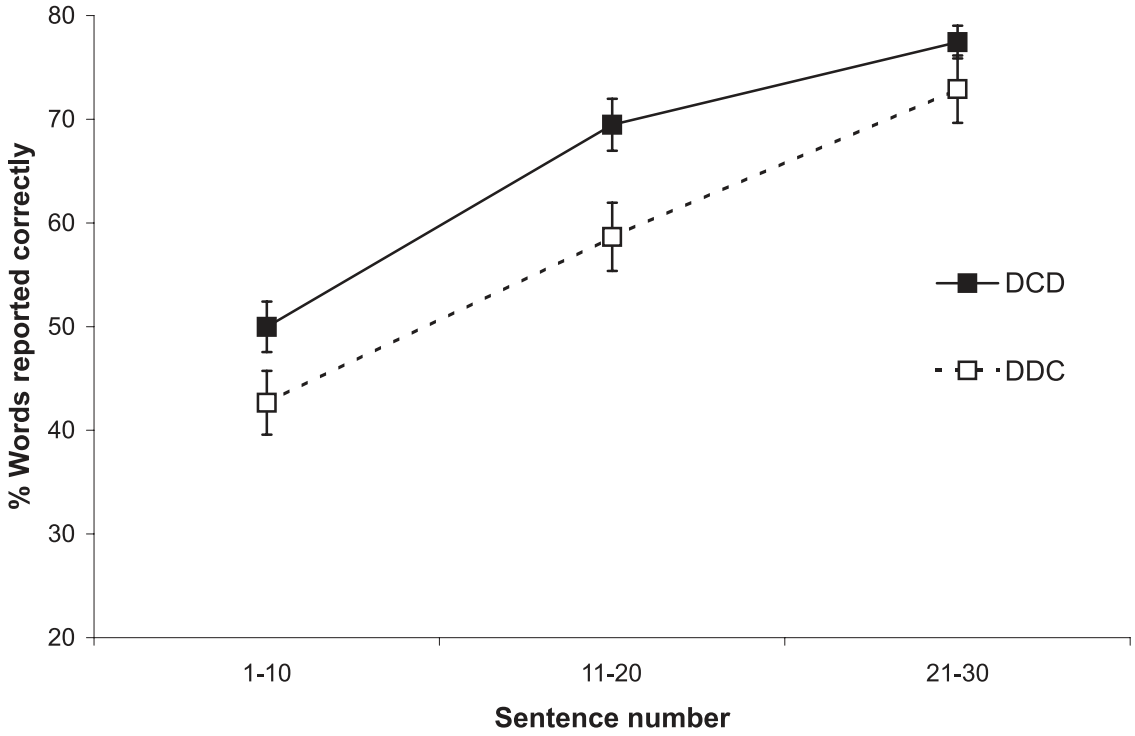

Figure 3. Report scores from Experiment 2 averaged over three groups of 10 sentences in two groups of participants. Error bars show plus or minus one standard error of the mean over participants. DCD = distorted–clear–distorted condition; DDC = distorted–distorted–clear condition.

group, $F_1(1, 40) = 125.60$, $MS = 6547$, $\eta^2 = .758$, $p < .001$; $F_2(1, 18) = 29.77$, $MS = 7211$, $\eta^2 = .623$, $p < .001$. Performance improves with increased exposure to noise-vocoded speech. More interesting, though, is that participants trained in the DCD condition (a condition likely to induce pop-out) performed significantly better than those trained in the DDC condition, a difference that was reliable by both participants and items, $F_1(1, 40) = 5.77$, $MS = 1609$, $\eta^2 = .126$, $p < .05$; $F_2(1, 18) = 7.79$, $MS = 1059$, $\eta^2 = .302$, $p < .05$. The interaction between block and condition was nonsignificant ($F_1 < 1$; $F_2 < 1$) both in analysis of performance in the first two blocks and when all three blocks were included in the analysis.[1]

A further point to note is that, in comparison with listeners in Experiment 1, listeners in Experiment 2 were markedly better at reporting noise-vocoded speech. For instance, report scores for the third block of sentences (Group C) exceeded 75% for participants in Experiment 2, whereas they failed to reach 60% in Experiment 1. This difference between performance in Experiments 1 and 2 is significant even where comparisons were made with participants in the poorer performing DDC condition, $t_1(31) = 2.41$, $\eta^2 = .158$, $p < .05$; $t_2(9) = 4.09$, $\eta^2 = .650$, $p < .01$. This confirms that learning of vocoded speech in Experiment 1 was not complete—further improvements in performance are possible—and suggests that hearing sentence repetitions benefits listeners in trying to understand vocoded speech even when these repetitions are not ordered in such a way to produce pop-out.

Discussion

The main result obtained from Experiment 2 is the demonstration that listeners who hear each vocoded sentence repeated after hearing the clear version are better able to report words from subsequent vocoded sentences than listeners who hear the same stimuli presented in a different order. Because both groups of participants reported vocoded sentences after only a single presentation and heard the same number of repetitions of vocoded and clear sentences subsequently, this difference in report score can be attributed to the order in which sentence repetitions were presented. Report scores for the two groups of participants did not differ for the very first test sentence, before listeners heard any sentence repetitions (mean report score for the first sentence: DCD = 4.5%, DDC = 5.1%) but differed between the two conditions for both the first and second block of sentences. This therefore indicates a difference in the rate of learning between the DCD and DDC conditions.

[1] Because the third test block involved a different group of sentences, analysis-combining results from all three blocks can only be conducted by participants and not by items. This analysis confirmed the pattern shown in the analysis of the first two sentence blocks with significant main effects of block, $F_1(2, 84) = 124.20$, $MS = 9301$, $\eta^2 = .747$, $p < .001$, and condition, $F_1(1, 42) = 5.61$, $MS = 1881$, $\eta^2 = .118$, $p < .05$, and no interaction between block and condition, $F_1(2, 84) = 1.45$, $MS = 109$, $\eta^2 = .033$, $p = .240$.

Only those participants in the DCD condition heard vocoded sentences when they knew the identity of all the words in that sentence, giving rise to pop-out. Although we cannot conclude that pop-out is directly responsible for the enhanced learning observed in the DCD condition, this result suggests that learning to understand vocoded speech is facilitated if listeners are able to use information on sentence content at the time that the vocoded sentence is repeated, a finding that would be consistent with top-down influences on learning.

Presentation of the clear sentence not only provides information on the content of the sentence but it also provides information on the acoustic form of clear speech. It is therefore possible that lower level mechanisms involved in extracting relevant features from the acoustic input are also responsible for at least some of the benefit that sentence repetition provides to the process of learning vocoded speech. An effect of repetition on lower level acoustic mechanisms might also explain the apparent benefit of hearing repetitions without top-down support, as indicated by improved performance by participants in the DDC condition compared with participants in Experiment 1.

Experiment 3: Written Feedback in Learning Vocoded Speech

To assess whether acoustic information is crucial for the advantage provided by presentation of a clear sentence, we tested whether provision of written feedback (which provides the content of a vocoded sentence without its acoustic form) supports learning as effectively as clear speech (which provides both the acoustic form and the content of the vocoded sentence). Thus, in Experiment 3 we compared performance in a condition in which a repetition of each vocoded sentence is preceded by written presentation of the sentence (i.e., distorted–written–distorted [DWD]) with the performance of volunteers tested in the DCD condition in Experiment 2. Should written presentation be as effective a source of feedback as presentation of the original spoken sentence, it would suggest the benefit provided by the clear sentence in DCD results from higher level information concerning the phonological or lexical content of the sentence and not its acoustic form. We also tested listeners in a distorted–distorted (DD) condition, in which each vocoded sentence is played twice without feedback. Performance on this condition defines a baseline level of performance against which any benefit of written feedback can be measured. If written feedback fails to provide any benefit to learning, then performance in the DWD condition would not differ from that in the DD condition. This null result would imply that the acoustic information, present in clear speech, is crucial to the advantage provided by the DCD condition.

Method

Participants. Twenty-four participants drawn from the same participant population were tested. All were native speakers of British English with no history of hearing or language impairment. None had taken part in the previous experiments or had prior knowledge of noise-vocoded speech.

Design and materials. The same three groups of 10 sentences (A, B, and C) were used as previously. Participants were randomly assigned to one of the two test conditions (DWD and DD) and to one of the two sentence orders (ABC, BAC). Data from these two conditions were compared with data from the DCD condition of Experiment 2.

Procedure. Participants in the DD condition were tested using the same equipment and procedure as in Experiments 1 and 2. In this condition, after writing down what they could understand of each vocoded sentence (and turning the page), participants were presented with a single repetition of the vocoded sentence, followed by a short tone indicating the beginning of the next trial.

In the DWD condition, participants were tested in single-participant booths using DMDX experimental software (Forster & Forster, 2003) running on a Windows 98 personal computer. Sentences were played using the computer sound card and the same amplifier and headphones as used previously. As before, participants in the DWD condition heard a single vocoded sentence, after which they had 25 s to write down as much as they could of the sentence that they had heard. A melodic sound instructed participants to stop writing and turn over the page in the answer book. A written version of each vocoded sentence was then displayed on the computer screen for 1.5 s and remained on the screen during the second presentation of the vocoded sentence. A warning tone then cued participants to expect the next vocoded sentence to report. Participants were supervised to ensure that sentence reports were only made in response to the first presentation of each vocoded sentence and not modified subsequently.

Results

Report scores were averaged as before and are shown (along with data from the DCD condition of Experiment 2) in Figure 4. Two-way ANOVAs were conducted to compare report scores for the first 20 items, for which item-specific differences in difficulty can be cancelled out. In analysis by participants, sentence block (first vs. second group of 10 sentences) was a repeated measures factor, and exposure condition (DCD vs. DWD vs. DD) was a nonrepeated factor. In the analysis by items, both factors (block and condition) were repeated measures factors. In both analyses, an additional dummy variable representing sentence group or sentence order was included, but effects of this factor will not be reported.

Results of these analyses showed a significant effect of block, $F_1(1, 41) = 102.00$, $MS = 4606$, $\eta^2 = .713$, $p < .001$; $F_2(1, 18) = 40.09$, $MS = 7105$, $\eta^2 = .690$, $p < .001$, confirming, once more, that improvements in report scores occur over the course of the experiment. The effect of condition was also significant, $F_1(2, 41) = 9.71$, $MS = 1766$, $\eta^2 = .212$, $p < .001$; $F_2(2, 36) = 18.88$, $MS = 2648$, $\eta^2 = .415$, $p < .001$, indicating that overall performance was affected by exposure condition. The interaction between block and condition was marginally significant by participants but not significant by items, $F_1(2, 41) = 2.52$, $MS = 114$, $\eta^2 = .109$, $p = .093$; $F_2(2, 36) = 1.56$, $MS = 232$, $\eta^2 = .080$, $p = .225$.[2]

Pairwise comparisons among the three conditions were conducted to determine the origin of the effect of exposure condition, using a Sidak correction for multiple comparisons (Toothaker,

[2] Results were essentially the same when data from all three blocks of sentences were considered in analysis by participants, with significant effects of block, $F_1(2, 82) = 194.60$, $MS = 8554$, $\eta^2 = .826$, $p < .001$, condition, $F_1(2, 41) = 9.06$, $MS = 2065$, $\eta^2 = .306$, $p < .001$, and a marginally significant interaction between these factors, $F_1(4, 82) = 2.22$, $MS = 98$, $\eta^2 = .098$, $p < .10$.

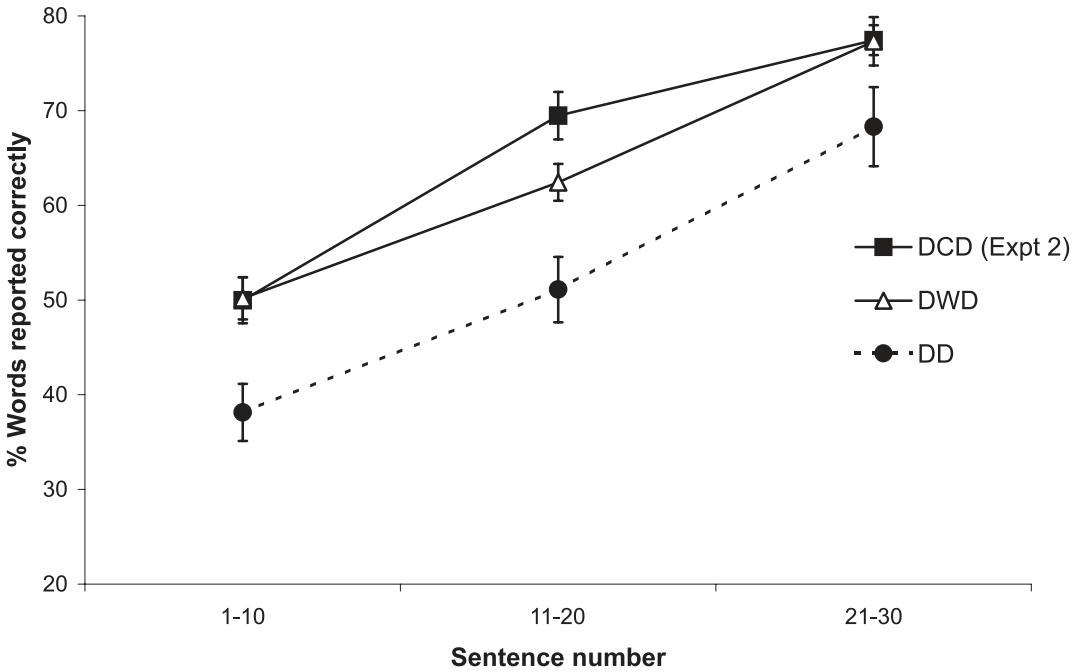

Figure 4. Report scores from Experiment 3 averaged over three groups of 10 sentences in three groups of participants (data from DCD in Experiment [Expt] 2 included for comparison). Error bars show plus or minus one standard error of the mean over participants. DCD = distorted–clear–distorted condition; DWD = distorted–written–distorted condition; DD = distorted–distorted condition.

1991). These comparisons indicated that performance was significantly worse in the DD condition compared with either the DCD condition ($p_1 < .001$; $p_2 < .001$) or the DWD condition ($p_1 < .05$; $p_2 < .001$), whereas performance did not significantly differ between the DCD and DWD conditions ($p_1 = .684$; $p_2 = .485$). This pattern of results indicates that the benefit provided by feedback is of equivalent magnitude with written and clear spoken presentation.[3]

Discussion

The results of Experiment 3 confirm the benefit that sentence content provides to listeners learning to understand vocoded speech. Presentation of either clear speech or a written form of each vocoded sentence prior to repetition of the vocoded sentence acts to increase performance on subsequent vocoded sentences. Thus, listeners' improved learning of noise-vocoded speech does not depend on access to low-level, acoustic information but can also be elicited by information concerning the phonological or lexical content of the sentence. Although listeners who receive a written sentence as feedback could generate a spoken form of the sentence by covertly reading it, they could not access the acoustic form of the original sentence. This result therefore demonstrates that the information supporting pop-out must be at a nonacoustic, phonological level or higher.

Experiment 4: Lexical Influences on Learning Vocoded Speech

The results of Experiments 2 and 3 demonstrate that the learning process responsible for improvements in the perception of vocoded speech is assisted by spoken or written feedback that provides information on the content of each sentence. This indicates a role for higher level, nonacoustic information in the learning process. However, on the basis of the results presented so far, we do not have any information on whether the crucial source of this higher level information is phonological (i.e., a representation of the speech sounds present in the sentence) or at a higher linguistic level (lexical, semantic, and/or syntactic). To address the possibility that phonological–phonotactic information benefits learning, we explored the effect of training listeners on vocoded nonword sentences—stimuli that conform to English phonology but that do not contain any familiar words.

[3] One anonymous reviewer suggested that comparison with a DDW condition (with written feedback following the repetition of distorted sentences) might be a more appropriate baseline than the DD condition included here. This reviewer argued that improved learning in the DWD condition might be modality-specific facilitation caused by presenting written stimuli in an experiment in which written responses are required. Although this is possible, it suggests two different explanations for the benefit observed for DCD and DWD training. We favor the simpler explanation that the benefit observed in both conditions results from providing sentence content prior to repetition of distorted speech. In support of this latter account, comparison of DDC and DD conditions in Experiments 2 and 3 failed to reach significance in the critical analysis by participants, $F_1(1, 29) = 1.97$, $MS = 598$, $\eta^2 = .064$, $p = .171$, although this comparison was significant by items, $F_2(1, 18) = 19.89$, $MS = 1403$, $\eta^2 = .525$, $p < .001$. By our account, we would predict no difference between results in the DD and DDW conditions.

Evidence from experiments investigating the perception of ambiguous fricatives suggests that lexical information plays an important role in perceptual learning of speech. In the study by Norris and colleagues (Norris et al., 2003), only those participants who heard ambiguous phonemes in familiar words showed subsequent changes to phoneme boundaries in a categorization task. This finding motivates the hypothesis that facilitation of learning observed in Experiments 2 and 3, when feedback on sentence content was provided before the repetition of the vocoded sentence, might reflect a similar dependence on lexical information. To test this hypothesis, we compared the performance of listeners trained on noise-vocoded English sentences with that of listeners who received an equivalent amount of exposure to vocoded sentences composed entirely of nonwords. Because we cannot expect participants to report nonword sentences, the design of this experiment was altered such that groups of listeners were exposed to either vocoded English or vocoded nonword sentences without any task. The effects of these training periods were assessed during a subsequent test period in which all listeners reported vocoded English sentences.

In training listeners with nonword sentences, one concern is that auditory–verbal STM capacity will be exceeded by nonword sequences that are more than a few syllables in length (Gathercole, Willis, Baddeley, & Emslie, 1994). If learning to understand vocoded speech depends on STM (for instance, in mapping distorted speech onto the equivalent sounds in clear speech), then nonword sentences might challenge learning for reasons other than a strict dependence on lexical information. In order to rule out STM as a critical limiting factor, we capitalized on the ability of written text to support STM—just as in everyday life people write down an unfamiliar name to support their limited STM for nonwords. We therefore tested a further group of participants following a training period during which vocoded nonword sentences were presented with written feedback (similar to the DWD training condition in Experiment 3, although without report). Presenting listeners with an orthographic transcription of the vocoded nonword sentences ensures that participants receive feedback on the phonological content of nonword sentences without placing any load on STM capacity. Requesting that participants read the visually presented sentence also helps ensure that they remain focused on the nonword sentences during training, potentially enhancing attention to the training stimuli.

Method

Participants. Twenty-four participants were randomly assigned to two of the three conditions (English or spoken nonword preexposure) and two sentence orders. An additional 12 participants were tested after training with nonword preexposure and visual feedback. All participants were drawn from the same participant population used previously, and none had any prior experience of noise-vocoded speech.

Design. Whereas listeners in the previous experiments were naive to noise-vocoded speech at the outset of testing, the participants in Experiment 4 were tested following a training period involving exposure to 20 vocoded sentences. One group of participants was trained with real English sentences. Two groups of participants were trained with vocoded versions of 20 nonword sentences created by replacing each word in the sentences with nonwords (see the *Materials* section below). Each vocoded sentence was presented twice during training. Between these two presentations, each sentence was repeated as clear speech (in spoken training conditions, this was equivalent to the DCD condition of Experiment 2 but without an initial report) or presented visually on a computer screen in front of the participants (for the written nonword training condition, as in the DWD condition of Experiment 3, but without report).

After this 20-sentence training period, all participants were tested on a further 20 English sentences (Groups A and C) as used in the previous experiments, presented either in the order AC (half the participants) or CA. During testing, participants were required to report as many words as possible after the first presentation of each vocoded sentence. They then heard each sentence clearly and then vocoded again (as in the DCD condition of Experiment 2). If the various training procedures used in this experiment yield the same learning as DCD in Experiment 2, then performance during the subsequent test will be equivalent to performance on the third block of sentences (Group C) from Experiment 2. We tested this, without a sentence-difficulty confound, in those participants in Experiment 4 who received the test order CA. If these training conditions confer no benefit, we would expect performance on the first block of test sentences to be like that of naive participants in Experiment 2: comparing performance in those listeners tested using the order AC with those who received Group A first in Experiment 2 will assess this. In this way, comparisons with trained and naive listeners can be made on the same groups of sentences.

Materials. Twenty English sentences (Groups D and E; see Appendix B, which is on the Web at http://dx.doi.org/10.1037/0096–3445.134.2.222.supp) were used to provide exposure to vocoded speech before testing. These were taken from the same set of sentences used in the earlier experiments (unambiguous sentences from Rodd et al., in press), matched on length, duration, rated naturalness, and imageability to Groups A, B, and C. Twenty nonword sentences were created from these sentences by replacing each word (both content and function words) with nonwords matched for length in syllables. For example, the real English sentence "The police returned to the museum" became the nonword sentence "Cho tekine garund pid ga sumeun."

These nonword sentences were recorded by the same native speaker of British English who produced the stimuli for the previous experiments. In recording the nonword sentences, we made every effort to match the nonword sentences to the English originals on rhythm and intonation by extensive rehearsals and by recording each pair of sentences (English and nonword equivalents) successively. Despite these efforts, however, there was a minor difference in speech rate, with the nonword sentences being acoustically longer than their English equivalents (mean durations: English = 2.0 s, nonword = 2.6 s), $t(19) = 10.77$, $\eta^2 = .859$, $p < .001$. Both English and nonword sentences were transformed using the same six-band vocoding process described previously and then recorded onto audio CD with instructions and warning tones as before.

Procedure. Participants in the two training conditions with spoken clear feedback were tested in groups of 2 or 3 in a quiet room using the same Sony CD player, QED amplifier, and headphones as used in Experiments 1 and 2. As before, the experiment was preceded by a four-item STM test and example (15-band) vocoded sentence. Following this introduction, participants were presented with either 20 vocoded English or 20 vocoded nonword sentences, presented as DCD with 2-s pauses between stimuli. Participants were instructed to listen attentively to all of the training materials but were not instructed to report words from any of these sentences.

Participants in the nonword training condition with written feedback were tested in booths using the same equipment as used in the DWD condition of Experiment 3. The training procedure was essentially the same as in the other two training conditions, except that the initial presentation of each vocoded nonword sentence was followed by visual presentation of the written form of the nonword sentence, which participants were instructed to read. The written nonword sentence continued to be displayed

on the computer screen during the subsequent repetition of the vocoded nonword sentence (with the same timings used in the feedback component of Experiment 3). Following these training trials, participants were tested on 20 English sentences (Groups A and C) presented for report with the same DCD procedure used in Experiment 2.

Results

Report scores for the 20 test sentences were averaged for participants pretrained in each condition and are shown in Figure 5. These data were entered into analyses by participants and items. In these ANOVAs, sentence block (Sentences 1–10 vs. 11–20) was a repeated factor; training condition was a nonrepeated factor in the analysis by participants and a repeated factor in the analysis by items. As before, a dummy variable in each analysis coded sentence group and sentence order in the analysis by participants and items.

These ANOVAs showed the expected effect of block, $F_1(1, 30) = 45.08$, $MS = 3385$, $\eta^2 = .600$, $p < .001$; $F_2(1, 18) = 19.42$, $MS = 5462$, $\eta^2 = .519$, $p < .001$, indicating significantly better performance on the second group of test sentences than the first. There was also a reliable main effect of training condition, indicating a significant difference in performance between the three groups exposed to different training conditions, $F_1(2, 30) = 6.61$, $MS = 1430$, $\eta^2 = .306$, $p < .01$; $F_2(2, 36) = 13.59$, $MS = 2384$, $\eta^2 = .430$, $p < .001$. Pairwise comparisons were conducted to determine the source of the main effect of training condition, Sidak corrected for multiple comparisons. These confirmed that participants trained on English sentences performed substantially better than those trained with nonword sentences and either spoken ($p_1 < .01$; $p_2 < .001$) or written ($p_1 < .05$; $p_2 < .001$) feedback. There was no significant difference in performance between the two groups of participants trained with nonword sentences ($p_1 = .814$; $p_2 = .647$).

There was a reliable interaction between block and condition, $F_1(2, 30) = 3.62$, $MS = 271.6$, $\eta^2 = .194$, $p < .05$; $F_2(2, 36) = 3.62$, $MS = 452.7$, $\eta^2 = .168$, $p < .05$. Assessment of simple effects is consistent with the source of this interaction being a ceiling effect for participants trained on English sentences. Performance showed a dramatic improvement between the two test blocks for participants who had been trained with nonword sentences and spoken, $t_1(11) = 4.24$, $\eta^2 = .621$, $p < .001$; $t_2(19) = 3.41$, $\eta^2 = .380$, $p < .01$, or written, $t_1(11) = 5.44$, $\eta^2 = .729$, $p < .001$; $t_2(19) = 3.92$, $\eta^2 = .446$, $p < .001$, feedback. However, the equivalent comparison for participants trained with 20 English sentences failed to reach significance, $t_1(11) = 1.69$, $\eta^2 = .206$, $p = .120$; $t_2(19) = 1.53$, $\eta^2 = .110$, $p = .114$, suggesting that report scores for this condition were near ceiling at the start of testing.

These results show that exposure to 20 English sentences (each presented three times as DCD) is sufficient for participants to subsequently report vocoded sentences with a high level of accuracy, even if the training procedure does not require participants to report each sentence. To determine whether the performance of the reporting task confers any benefit over that of passive training, we compared report scores for Group C sentences in the DCD condition of Experiment 2 with report scores for those participants who received Group C immediately after exposure to (but not report from) 20 English sentences in the current experiment. This com-

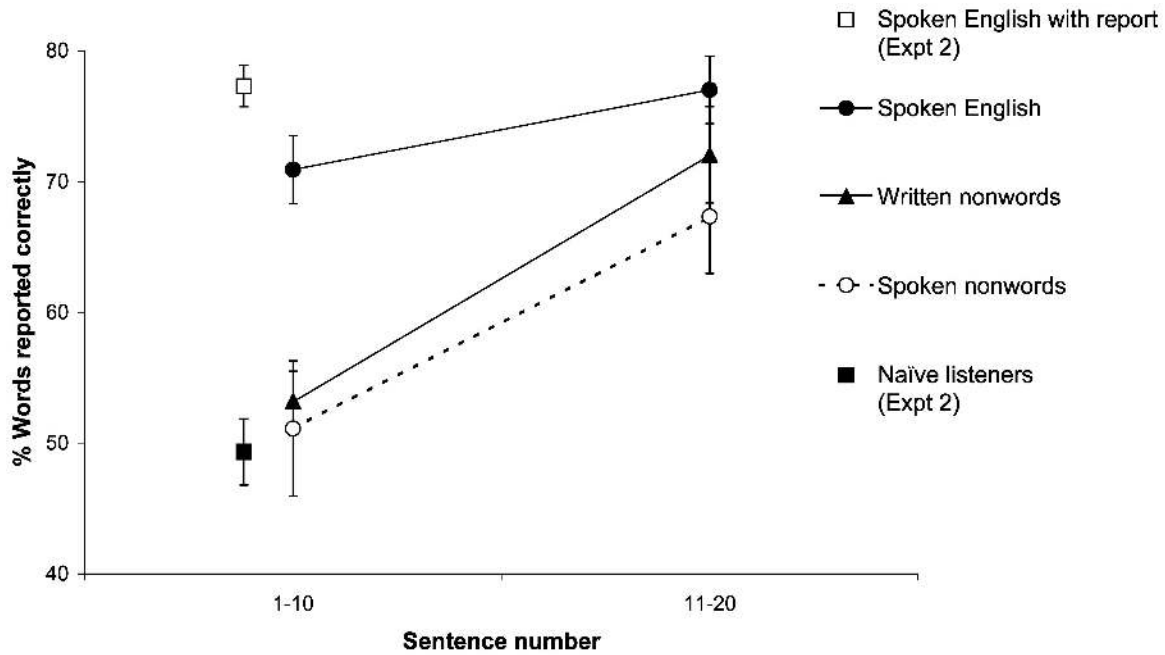

Figure 5. Report scores from Experiment 4 averaged over two groups of 10 sentences. Data from the first group of sentences for naive and trained listeners in Experiment (Expt) 2 included for comparison. Error bars show plus or minus one standard error of the mean over participants.

parison showed a marginally significant difference by participants, though this was not significant by items, $t_1(27) = 1.71$, $\eta^2 = .097$, $p = .099$; $t_2(9) < 1$.[4] This result suggests that the process of learning to understand vocoded sentences is not reliant on participants performing an active task; passive listening is sufficient for learning to take place, at least if sentences contain real English words.

Results showed better report scores for volunteers trained on English vocoded sentences compared with those trained with nonword sentences. This difference clearly demonstrates that lexical information in sentences benefits listeners learning to understand noise-vocoded speech. However, it is still possible that some benefit can be derived from training with nonword sentences. To assess whether training with nonword sentences confers any benefit at all, we compared performance with that of naive listeners initially tested on Group A in Experiment 2. This comparison shows no significant difference between the performance of naive listeners and those trained with 20 vocoded nonword sentences with spoken feedback, $t_1(17) = 1.26$, $\eta^2 = .085$, $p = .225$; $t_2(9) < 1$, or written feedback, $t_1(17) = 1.23$, $\eta^2 = .081$, $p = .237$; $t_2(9) = 1.45$, $\eta^2 = .188$, $p = .182$, even where these two conditions are combined to increase statistical power, $t_1(23) = 1.64$, $\eta^2 = .104$, $p = .116$; $t_2(9) < 1$.[5] Participants in the nonword training conditions of Experiment 4 have heard numerous presentations of vocoded speech, accompanied by clear speech or written feedback, yet their report scores are indistinguishable from listeners who are entirely naive to noise-vocoded speech.

Discussion

The results of Experiment 4 demonstrate a critical condition that is required for listeners to learn from exposure to vocoded speech: They should be hearing sentences containing real words. Listeners exposed to 20 vocoded nonword sentences, along with clear speech equivalents or written feedback, are no better at reporting English vocoded sentences than listeners who are entirely naive to vocoded speech. The lack of any difference between nonword training conditions with spoken or written feedback suggests that this failure of training with nonword sentences is not merely due to limitations on STM capacity. In Experiment 3, written feedback assisted learning just as effectively as spoken feedback (i.e., there was no difference between DWD and DCD conditions) and thus could have been of benefit in Experiment 4, particularly because STM was not being challenged by the demands of remembering nonword sentences. Nevertheless, subsequent report scores for English sentences were still no better than naive performance.

This result suggests that higher level lexical, semantic, and/or syntactic information in normal sentences plays a critical role in learning to understand noise-vocoded speech. However, this finding alone does not determine what information in sentences drives perceptual learning of noise-vocoded speech. To examine whether it is the presence of lexical items themselves, or whether it is semantic or syntactic information at a level of representation beyond single words that is crucial for learning, we further manipulated the content of the training sentences in Experiment 5.

Experiment 5: Semantic and Syntactic Content During Training

All of the training conditions that produced successful learning in Experiments 1 to 4 used vocoded speech that contained both sentence-level meaning and grammatical structure in addition to real words. It is possible that semantic or syntactic information, a level of representation that is beyond single words, might be crucial for learning. We tested the possibility that learning depends on sentence-level meaning by training listeners with *syntactic prose* sentences (cf. Marslen-Wilson, 1985): These are sentences in which content words are randomly replaced so as to create prose that is grammatical and composed of real words but lacks sentence-level meaning (see Table 1). If training with syntactic prose produces equivalent benefit to training with real English, then we can conclude that sentence-level meaning is not critical for learning to take place.

We tested the importance of grammatical structure by training with Jabberwocky sentences. These are sentences containing real English function words but in which content words are replaced with nonwords (see Table 1; cf. Friederici, Meyer, & von Cramon, 2000; Marslen-Wilson, 1985). Prior experiments have shown that Jabberwocky sentences are as effective as real English in training listeners to understand time-compressed speech (Altmann & Young, 1993). Should a similar result be obtained here, it would suggest that syntactic information (largely carried by function words) is sufficient for learning to take place. In this experiment, we compared groups trained with these two new sentence types with three other groups: listeners trained with real English, a group of naive listeners, and a group of listeners trained with nonword sentences. This last condition was included to replicate the comparison between training with English and nonword sentences from Experiment 4.

Method

Participants. Eighty participants from the same participant population sampled in the previous experiments were tested. None of the participants had any previous experience with noise-vocoded speech. Recruitment and testing proceeded in two stages, with 40 participants tested in each stage. Within each stage, participants were randomly assigned to different test conditions, though no naive participants were tested initially, and no

[4] This comparison between performance on Group C sentences only includes data from 6 of the 12 participants tested following training on English. If we ignore variance in report scores produced by different groups of sentences, we can include additional data in this comparison, revealing a significant difference between scores for participants who report from all sentences and those exposed to vocoded sentence without report, $t_1(33) = 2.27$, $\eta^2 = .134$, $p < .05$. However, we cannot conduct an items analysis to rule out the possibility that this result arises from differences in sentence difficulty.

[5] We can include more data in this comparison if we combine data from the different sentence groups in analysis by participants, though this comparison will be vulnerable to false positives caused by differences in sentence difficulty. Nonetheless, this analysis still shows no significant difference in test performance between naive listeners and those trained with 20 vocoded nonword sentences including written or spoken feedback, $t_1(45) = 0.56$, $p > .10$.

Table 1
Training Conditions and Example Sentences From Experiment 5

Training condition	Example sentence
Nonword	Cho tekeen garund pid ga sumeeun.
Jabberwocky	The tekeen garund to the sumeeun.
Syntactic prose	The effect supposed to the consumer.
Real English	The police returned to the museum.

participants were trained with real or syntactic prose in the second stage. Comparisons between participants in training conditions common to the first and second recruiting stage revealed no significant difference in performance.

Design and materials. The design was like that of Experiment 4, with a period of training with 20 DCD items without report followed by two 10-item blocks of real English sentences that participants reported after the first vocoded presentation. In addition to the two training conditions of Experiment 4, two new conditions were created from Groups D and E used in Experiment 4. Jabberwocky sentences, containing English function words and nonsense content words, were created by taking the nonword versions of Groups D and E and reinstating the function words and inflectional endings from the original sentences. Syntactic prose sentences were created by replacing randomly chosen content words from the English version of D and E with semantically unrelated words matched for length (in phonemes and syllables) and frequency (taken from the CELEX lemma database; Baayen, Piepenbrock, & Gulikers, 1995). Examples of each of these conditions are shown in Table 1. A full list of sentences used in each training condition is provided in Appendixes B and C (which are on the Web at http://dx.doi.org/10.1037/0096–3445.134.2.222.supp).

The same speaker who recorded the original English sentences and the nonword sentences in Experiment 4 recorded the training sentences in the syntactic prose and Jabberwocky conditions. The speaker once again attempted to maintain the same prosody as used in the original English sentences, although differences in speech rate were still observed (mean sentence durations: English = 2.0 s, syntactic prose = 2.3 s, Jabberwocky = 2.4 s, nonword = 2.6 s), $F(3, 57) = 44.21$, $MS = 1$, $\eta^2 = .699$, $p < .001$. All training sentences were vocoded using Praat software and recorded onto audio CD with instructions and warning tones as before.

Procedure. All participants received a four-item memory test before the start of the experiment (on which all scored 100%) and were played a single 15-band vocoded sentence as an example. In each of the four trained groups (i.e., in all except the naive condition), listeners first heard the 20 training items as DCD, without report, followed by testing on 20 English DCD items. For these items, they were required to write down what they had understood from the sentence after the first vocoded presentation. A group of naive listeners was tested directly without any training (as in Experiment 2). In two groups, the conditions were identical to Experiment 4: Groups D and E, either English or nonword sentences. Another group was trained with Jabberwocky sentences, and the fourth group was trained with syntactic prose. As before, listeners were not expected to report the sentences used during training but were instructed to listen attentively. The test session comprised Sentence Groups B and C, presented in counterbalanced order, though Sentence B7 was replaced with Sentence A7 for this study.

Results

Report scores for the 20 test sentences were averaged over participants in each condition and are shown in Figure 6. These data were also entered into repeated measures ANOVAs, con-

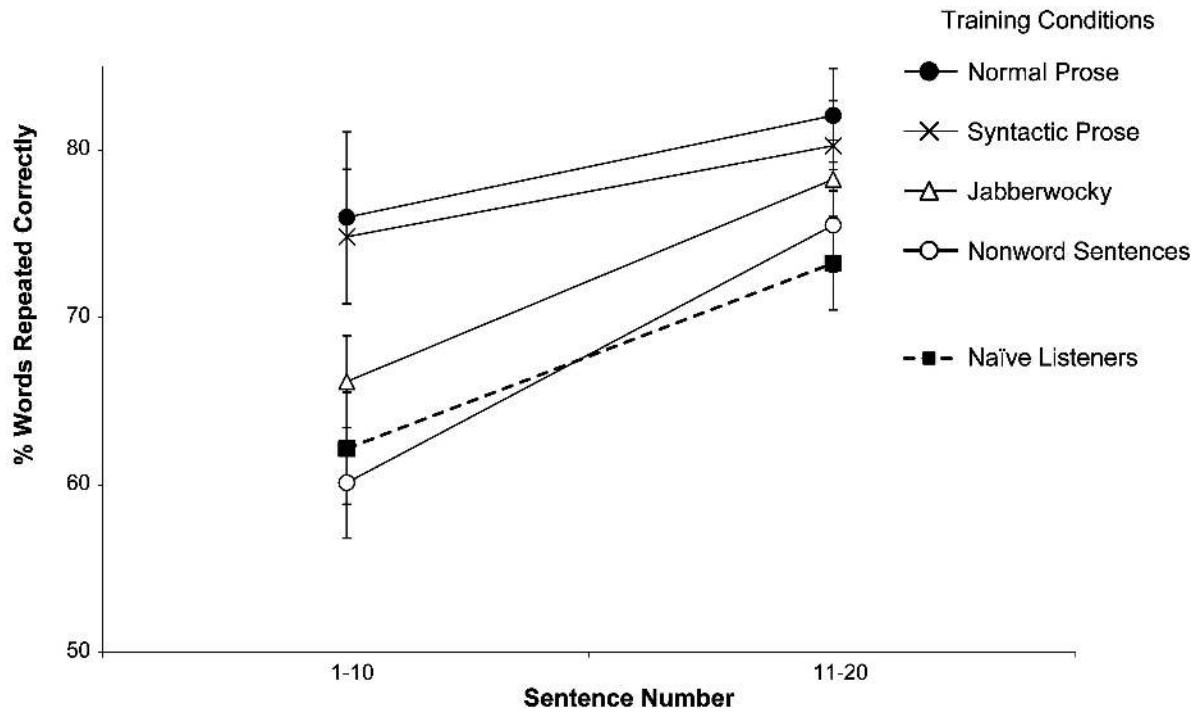

Figure 6. Report scores from participants pretrained with different types of sentences and naive listeners from Experiment 5. Error bars show plus or minus one standard error of the mean over participants.

ducted both by participants (F_1) and items (F_2). Sentence block was a repeated factor in both analyses; condition was a nonrepeated factor in the analysis by participants and a repeated factor in the analysis by items. As previously, a dummy variable was used to code sentence group and order throughout these analyses.

Once again, these analyses revealed a significant effect of block, $F_1(1, 70) = 44.45$, $MS = 3435$, $\eta^2 = .388$, $p < .001$; $F_2(1, 18) = 30.21$, $MS = 4775$, $\eta^2 = .627$, $p < .001$, confirming that performance on the second set of 10 test sentences was significantly better than on the first. There was also a significant main effect of condition, $F_1(4, 70) = 4.30$, $MS = 752$, $\eta^2 = .197$, $p < .01$; $F_2(4, 72) = 14.35$, $MS = 1130$, $\eta^2 = .444$, $p < .001$, indicating that the type and/or presence of training had a significant effect on performance. The interaction between training condition and block was nonsignificant, $F_1(4, 70) = 1.51$, $MS = 117$, $p = .209$; $F_2(4, 72) = 1.78$, $MS = 163$, $p = .143$.

To analyze the effect of condition further, we performed post hoc pairwise comparisons using a Sidak correction for multiple comparisons. These showed that the performance of listeners trained with real English sentences was significantly better than that of listeners trained with nonwords ($p_1 < .05$, $p_2 < .001$) and naive listeners ($p_1 < .05$, $p_2 < .001$), and performance in the nonword and naive conditions did not differ ($p_1 = p_2 = 1.00$). These results replicate the findings of Experiment 4. Performance in the syntactic prose condition was also better than performance in both the nonword ($p_1 = .087$, $p_2 < .05$) and naive ($p_1 = .080$, $p_2 < .05$) conditions, although these differences were only marginally significant by participants after correction for multiple comparisons. As is apparent from Figure 6, performance in the English and syntactic prose conditions did not differ ($p_1 = 1.00$, $p_2 = .995$), indicating that disrupting the semantic content of sentences did not significantly reduce the efficacy of training.

Results of comparisons involving the Jabberwocky condition were equivocal, with significant differences in analyses by items but no reliable differences in analyses by participants. In view of the nonsignificant differences observed in pairwise comparisons of naive and nonword-trained listeners, and English and syntactic-prose-trained listeners, we combined each of these pairs of conditions to boost power for comparisons with the Jabberwocky condition. Exploratory uncorrected comparisons showed that the group trained with English or syntactic prose performed significantly better than the group trained with Jabberwocky sentences ($p_1 < .05$; $p_2 < .01$), and the group comprising naive listeners and those trained with nonword sentences performed marginally more poorly than the Jabberwocky-trained group ($p_1 = .069$, $p_2 < .01$).

Discussion

The pattern of results across the four training conditions confirms and extends the results of Experiment 4 and indicates that lexical information, not syntactic or sentence-level semantic information, drives learning. Syntactic prose (without sentence-level meaning) was as effective at producing learning as semantically coherent English sentences, and nonword sentences were entirely ineffective, producing no improvement over naive performance. Function words are a core element of syntactic structure, but Jabberwocky sentences in which function words are preserved were significantly less effective than training with sentences composed exclusively of words: This finding suggests that syntactic content alone is not sufficient to drive learning. Jabberwocky sentences were somewhat more effective than no training or training with nonword sentences, most probably because of the presence of some real words in these sentences.

General Discussion

In a series of five experiments, we investigated the processes by which listeners learn to recognize a form of artificially distorted speech. Noise-vocoded speech, as used in our studies, is initially unintelligible, with participants able to report few if any words from the first noise-vocoded sentence that they hear. However, over the course of 20-min experiments, involving exposure to 30 or 40 vocoded sentences, report scores increase rapidly: Listeners are able to report most of the words in each sentence by the end of the experiment. This dramatic change in participants' comprehension reflects the operation of powerful mechanisms that substantially alter the perception of distorted speech.

What processes are responsible for this change in the way that vocoded speech is perceived? In order to address this question, we must first establish what it is that participants are learning. We can conclude that listeners are learning a general property of vocoded speech rather than learning the sound of individual vocoded words. Performance improves not only on those words that had been heard in previous trials but also on words that had not already been heard in vocoded form. For instance, of the words used in the Group C sentences (presented last in Experiments 1–3), only 46% occurred in the Group A and B sentences. Even in Experiment 1, when listeners did not receive any feedback on the content of previous vocoded sentences (and learning was least effective), report scores for Group C sentences were significantly higher than 46%, $t(11) = 2.34$, $\eta^2 = .332$, $p < .05$, and 11 out of 12 participants reported more than 46% of words correctly. A recent study in which participants reported isolated vocoded words provides further confirmation that perceptual learning does generalize to words not previously heard in vocoded form (Hervais-Adelman, Carlyon, Davis, & Johnsrude, 2004).

Our observation of improved report score on words that listeners have not previously heard in vocoded form necessarily implies that training produces changes in perceptual processing at a level of the comprehension system at which parts of words (and not whole words) are represented. In a hierarchically organized model of speech perception, we would therefore conclude that learning must be altering perceptual processing at a prelexical level.

The present results do not allow us to draw any more precise conclusions concerning the prelexical representations that change during training. For instance, it is unclear whether the learning process affects representations that are organized in terms of phonetic features, phonemes, or syllables as proposed in various models of speech perception. However, transfer of learning from trained to untrained items could be used in studies to investigate the nature of the prelexical representations that are altered by learning. For example, if training with only a subset of phonemes produces good performance on untrained phonemes, this would suggest that the locus of learning was subphonemic. Other levels of representation that are assumed to mediate between the incoming speech signal and higher level processes (such as syllables or

acoustic–phonetic features) could similarly be tested. Research testing for generalization between different forms of vocoded speech (carrier signals, frequency ranges, etc.) could similarly determine whether learning involves attending more strongly to those acoustic features (e.g., amplitude envelope) that are preserved in vocoded speech (an attentional weighting process; cf. Goldstone, 1998). This method for using generalization of perceptual learning to assess intermediate levels of representation in speech perception is reminiscent of the psychoanatomical method described for vision by Julesz (1971). For instance, testing whether perceptual learning of visual orientation detection generalizes from a trained to an untrained eye, or between different retinotopic regions, provides a means by which vision scientists can assess the level of the visual system at which learning occurs (see Ahissar & Hochstein, 2004, for further discussion).

Our conclusion, that learning arises at a prelexical level, points to an apparent inconsistency: Although learning produces changes to perceptual processing at a prelexical level, our experiments also provide clear evidence that higher level lexical information is crucial for learning. In Experiment 2, learning was enhanced for listeners who knew what they were hearing when the vocoded sentence was repeated. This benefit results from knowledge of sentence content (and not acoustic form), because an equivalent advantage was observed in Experiment 3 using written feedback. Confirmation that lexical processes are involved in learning comes from the results of Experiments 4 and 5, which show that listeners trained on vocoded nonword sentences are no better at reporting English vocoded sentences than naive listeners. Exposure to noise-vocoded sentences without familiar words does not permit learning. However, as shown in Experiment 5, semantic content at a supralexical level is unnecessary for learning, nor is syntactic information in the absence of content words (as in Jabberwocky sentences) a fully effective training stimulus.

Our results therefore demonstrate that a top-down, lexically driven mechanism is involved in perceptual learning of noise-vocoded speech. The learning process is reliant on information at the lexical level, but this information is used to make alterations to perceptual processes at a prelexical level. The search for top-down processes in speech perception (long seen as a holy grail; Norris et al., 2000) now seems to have been resolved through our demonstration (and other findings, e.g., Norris et al., 2003) of top-down, lexically driven learning. These findings have important implications for models of spoken language comprehension, in which information flow from lexical to prelexical processes has been the subject of considerable recent debate (e.g., Norris et al., 2000). Similar arguments for top-down influences in vision can be found in the reverse-hierarchy theory of Hochstein and Ahissar (Ahissar & Hochstein, 2004; Hochstein & Ahissar, 2002). Consistent with our observations for speech, these authors proposed that information used for perceptual learning originates in higher level visual areas (representing objects or complex combinations of features), which feed information back to lower level areas tuning representations of simpler visual features.

Implications for Models of Speech Perception

Our findings provide evidence by which to assess an important distinction common to many models of spoken language comprehension (as in other hierarchically organized systems), specifically, whether later stages of processing can influence earlier stages. Models of speech perception make different predictions regarding the influence of word recognition on the activation of sublexical representations of the speech input. Models range from being strongly interactive (i.e., models in which sublexical processes are influenced by top-down feedback from later stages, e.g., TRACE; McClelland & Elman, 1986) to models in which top-down feedback plays little or no role in perception (e.g., Shortlist/MERGE; Norris, 1994; Norris et al., 2000). Other intermediate positions have also been explored (Gaskell & Marslen-Wilson, 1997).

Our characterization of the process of learning to understand noise-vocoded speech indicates that lexical information is necessarily involved in tuning sublexical processes. Our results therefore imply that some form of top-down feedback is in operation. This result at first sight seems most compatible with strongly interactive accounts such as TRACE. However, TRACE is a network model in which localist units (representing features, phonemes, and words) are hard wired, without any learning mechanism available to alter connectivity. Although learning mechanisms are possible in localist systems (see Page, 2000), no model of spoken language comprehension has thus far been proposed that learns to develop localist internal representations such as are incorporated in TRACE.

Those models of speech perception that most readily incorporate learning are those constructed using recurrent neural networks such as the distributed cohort model (DCM; Gaskell & Marslen-Wilson, 1997). Recurrent network models (such as DCM) implement different types of information flow in training the network and in online processing. In DCM, information flows in a bottom-up fashion during testing, whereas training is an interactive process, involving both bottom-up and top-down processing. These networks use a supervised learning algorithm (back propagation) to acquire and maintain the ability to recognize spoken words (see Davis, 2003, for an evaluation of developmental assumptions; see Norris, 1993; Norris et al., 2000, for further discussion). The results of the present experiments suggest a further role for supervised learning in altering perceptual-level processes so as to compensate for systematically distorted input. One informal description of this learning process might be as follows: Once words have been recognized, the activation of stored lexical representations provides information on the speech sounds that must have been present in the input. Top-down mechanisms, supervised using this information, can then retune lower level perceptual processes to output the required sublexical units (whether these are specified as features, phonemes, etc.). This top-down supervised learning process provides a clear explanation of the absence of learning for nonword sentences (because there can be no target representation for a nonword) and for the benefit of training with known sentences (because this provides a clearer target representation). A similar top-down mechanism has been proposed for learning to recognize ambiguous phonemes (Norris et al., 2003).

On the basis of the current set of experiments, we contend that this top-down learning mechanism operates rapidly and automatically. The speed with which improvements in performance occur is striking—our experiments included just 30 or 40 trials and lasted around 20 min. The fact that feedback on sentence content

is most effective if it is available at the same time as distorted speech is presented (i.e., the difference between DCD and DDC in Experiment 2) suggests that learning depends on information that is available while the second vocoded sentence is being presented. Indeed, comparisons between DDC and DD training were nonsignificant (see Footnote 3), suggesting that feedback after the presentation of the vocoded sentence does not assist learning. Furthermore, listeners in Experiment 4 learned as much from passive exposure as listeners in Experiment 2 learned from active report. This suggests that perceptual learning of speech may occur automatically, at least when attentional resources are not explicitly deployed elsewhere.

One intriguing aspect of our results is that those situations that produce the most effective learning in these experiments also produce pop-out, a dramatically clearer percept of an otherwise relatively unintelligible vocoded sentence. Although we cannot be certain that pop-out and perceptual learning reflect operation of the same top-down processes, it seems most parsimonious to believe that the same mechanism of top-down information flow both facilitates learning and alters perception simultaneously.

The rapid learning that we have observed is not typically associated with networks trained using back propagation, which often exhibit a trade-off between the speed of new learning and the stability of previously acquired knowledge (French, 1999; McClelland, McNaughton, & O'Reilly, 1995; Page, 2000). However, other supervised learning algorithms exist, including those in which sparse or localist representations mediate between the speech input and lexical output, and these might be capable of simulating a more rapid learning process. Furthermore, because longer term effects of learning were not assessed, we cannot be certain that a form of offline consolidation (more consistent with back-propagation style learning) is not also present for learning vocoded speech (see McClelland et al., 1995). Recent evidence for offline consolidation has been obtained from studies of word learning (Gaskell & Dumay, 2003) and learning to understand synthesized speech (Fenn, Nusbaum, & Margoliash, 2003). The long-term consequences of our training procedures on perception of noise-vocoded speech will be a topic of future work.

Our results clearly point to a top-down perceptual learning process driven by lexical and not supralexical (semantic–syntactic) information. This is not to say that supralexical top-down influences may not exist in other circumstances. For example, ambiguities created by a phonological process (assimilation) may be resolved in a similar fashion as lexical ambiguities (Gaskell & Marslen-Wilson, 2001). However, other research has shown that higher level information (such as sentential context) has a qualitatively different influence on phoneme perception than lexical information (Connine & Clifton, 1987; Samuel, 1981). Supralexical influences on phoneme identification and restoration that have been observed may reflect postperceptual strategic processes.

Comparison With Other Forms of Degraded Speech

Having explored the cognitive processes involved in learning to understand noise-vocoded speech, we can ask whether other forms of distorted speech are learned the same way. Speech comprehension can be challenged by acoustic manipulations that affect spectral (Faulkner et al., 2000; Remez et al., 1981) or temporal (Mehler et al., 1993; Saberi & Perrott, 1999) properties of speech, or more simply by masking the speech signal with noise (G. A. Miller, Heise, & Lichten, 1951) or other speech sources (Cherry, 1953). In a more naturalistic context, the considerable variability that exists between speakers presents a considerable challenge to the perceptual system (Nusbaum & Magnuson, 1997), and perceptual learning appears to play a role in compensating for this variation (Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994).

Is learning of these various forms of degraded speech similar to learning of noise-vocoded speech? Are the top-down learning processes that we have documented unique to certain forms of degraded speech or general to all? For manipulations that irretrievably mask information in the speech signal (such as speech in background noise), low-level mechanisms that permit the separation of speech and noise will determine the intelligibility of speech (Carhart, Tillman, & Johnson, 1967; Cherry, 1953). However, there is little evidence to suggest that the spectral and temporal mechanisms responsible for separating speech and noise (Carhart et al., 1967; Dubno, Ahlstrom, & Horwitz, 2002) can be modified by experience to produce the rapid and dramatic improvements in intelligibility that we have observed in our experiments (Peelle & Wingfield, 2003).

Perhaps it is only those forms of distortion that retain critical information in an altered form that are amenable to perceptual learning? For instance, comprehension of time-compressed speech increases over the first 10 min of listening (Foulke & Sticht, 1969; Voor & Miller, 1965). Experiments have demonstrated that this adaptation is independent of acoustic properties of the stimulus (such as speaker identity or amount of compression; Dupoux & Green, 1997), though in contrast to our Experiment 5, training with time-compressed Jabberwocky appears as effective as training with real sentences (Altmann & Young, 1993; Mehler et al., 1993). Furthermore, some adaptation is observed even if listeners are trained with a foreign language that they do not understand (such as training English-speaking listeners with time-compressed Dutch; Pallier et al., 1998). In subsequent investigations, cross-language transfer appears to be confined to language pairs that share rhythmic structure and other phonological properties (e.g., vowel inventory and lexical stress) such as Spanish and Catalan or English and Dutch (Pallier et al., 1998; Sebastian-Galles, Dupoux, Costa, & Mehler, 2000). These results suggest that learning of time-compressed speech depends on phonological information rather than the lexical information that is crucial for vocoded speech. However, direct comparisons of these two forms of distortion would be needed to confirm this dissociation.

A form of artificial distortion in speech that may be similar to vocoded speech is generated by resynthesizing speech formants as sinusoidal tones. Listeners' perception of sine-wave speech is similarly affected by information concerning the content of the distorted sentence (Remez et al., 1981). Experiments that use sine-wave speech (e.g., Best, Studdert-Kennedy, Manuel, & Rubin-Spitz, 1989; Remez, Rubin, Nygaard, & Howell, 1987) include training with feedback (similar to our DCD manipulation) to enable participants to report a speech percept for sine-wave stimuli. However, comparisons of the efficiency of different procedures for learning to understand sine-wave speech have not been conducted.

In addition to investigations of artificially distorted speech, research has also explored more natural forms of variation. Notably, Nygaard and colleagues (Nygaard & Pisoni, 1998; Nygaard et al., 1994) demonstrated that familiarity with a speaker's voice (produced by several days training on voice identification) results in improved speech identification in noise. Research using heavily accented speech (produced by nonnative speakers) similarly demonstrates improved intelligibility after 5 days of training (Weill, 2001), with some transfer of learning to other speakers with a similar accent. For other (perhaps more familiar) accents, a more rapid training effect can be observed (Clarke & Garrett, 2004). Improvements in intelligibility can also follow training on poor-quality computer-generated speech (Schwab, Nusbaum, & Pisoni, 1985). It is striking that learning this form of speech is consolidated by an overnight period of sleep (Fenn et al., 2003). Further research might therefore explore whether a form of lexically driven perceptual learning may be involved for these other forms of distortion.

Implications for Learning Cochlear Implant Transduced Speech

The experiments reported here used a form of distorted speech intended to capture properties of speech transduced by a cochlear implant, through reductions in spectral detail and changes to temporal fine structure. Our results might suggest learning processes that are engaged when individuals with a hearing loss receive a cochlear implant. However, there remain some important details of the noise-vocoded simulation that do not adequately capture the input provided by the current generation of cochlear implants. One possible reflection of these differences is that the learning we observed is considerably more rapid than the lengthy period of rehabilitation typically required by cochlear implant users. It is possible that differences between our simulations and cochlear implant-transduced speech can explain these differences in the time course of training.

In the vocoded simulations used here, we have provided six separable bands of temporal information. Prior work suggests that listeners presented with vocoded speech containing fewer bands might require more extensive training (see Shannon et al., 1995, 1998). Although current cochlear implants typically contain many more electrodes (for instance, the CI24M cochlear implant [Cochlear Corporation, Lane Cove, New South Wales, Australia] has 22 intracochlear electrodes), it is never the case that all of the electrodes can be used to discriminably stimulate the auditory nerve. Typical cochlear implant processing schemes, such as the continuous interleaved sampling strategy (Wilson et al., 1991), provide approximately six separable bands of information and are therefore broadly comparable with the stimuli used in our studies.

A second difference between our simulations and cochlear implant-processed speech is the form of the carrier of the fine-structure information. Our vocoded simulations used a noise carrier, whereas the most commonly used implant processing scheme applies pulse trains to each electrode (Wilson et al., 1991). It has been demonstrated that simulations in which amplitude modulations are superimposed onto pulse-train (rather than noise) carriers may provide a more accurate simulation of the input received by cochlear implant users, especially where pitch perception is concerned (Carlyon, van Wieringen, Long, Deeks, & Wouters, 2002). Although we do not anticipate differences in how noise- and pulse-train vocoded speech is learned, further investigations and comparisons may prove valuable.

A final, and perhaps more crucial, difference between our simulations and speech transduced by a cochlear implant is that the modulated bands of noise stimulate the same frequency region as analyzed from the original speech in our simulations. Because of physical difficulties in inserting an electrode array into the regions of the cochlea that respond to lower frequency sounds, this direct frequency mapping is not typically achieved with a cochlear implant. Instead, cochlear implant users often receive speech information transposed to a higher frequency region of the cochlea (see Dorman et al., 1997; Rosen et al., 1999, for discussion). Experimental investigations of vocoded simulations that mimic these frequency shifts typically show a much slower learning process than is typical for nonshifted simulations (Fu & Galvin, 2003; Rosen et al., 1999), suggesting that these frequency shifts are an important source of additional difficulty for cochlear implant users. Converging evidence from the perception of speech produced by divers in a helium-rich environment (Belcher & Hatlestad, 1983; Morrow, 1971) supports the proposal that pitch shifting of speech formants creates a substantial additional obstacle to comprehension (even in the absence of any loss of spectral detail). Further experiments using pitch-shifted vocoded simulations may validate the role of feedback for stimuli that require a longer period of training to achieve successful comprehension.

Although perceptual learning of speech by cochlear implant users is therefore different from the learning problem faced by our normally hearing listeners, the factors that we have discovered to facilitate learning in our simulations may still apply to cochlear implant users. In fact, it would be surprising if lower level mechanisms alone were sufficient in more demanding situations. We might expect that more challenging forms of distorted speech (such as cochlear implant-transduced speech) might be even more dependent on higher level processes. The optimal conditions for learning established in this article, emphasizing a role for lexical information and for the provision of information on speech content, may therefore be relevant to the development of rehabilitation programs for cochlear implant users. However, the complex neurobiology of deafness and cochlear implantation cannot ever be entirely simulated, and future research with hearing-impaired populations and cochlear implant users will be required to determine whether our results can indeed inform clinical practice.

In summary, our studies illustrate the processes by which listeners adjust to unusual or unfamiliar-sounding speech. We have demonstrated that learning to understand noise-vocoded speech produces changes to prelexical representations but also requires top-down, lexical feedback. This is an important step toward understanding the cognitive mechanisms that allow listeners to understand speech in spite of considerable variability in the form of spoken input. This lexically driven learning process likely plays a role in many different situations in which the perceptual system is challenged by distorted or degraded speech input. More generally, the results of these studies point to a role for top-down processes in tuning the perceptual system to optimally perceive subsequent input.

References

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Science, 8,* 457–464.

Altmann, G., & Young, D. (1993). Factors affecting adaptation to time-compressed speech. *EUROSPEECH'93,* 333–336.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* (Version 2.5) [CD-ROM]. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Bailey, P. J., & Summerfield, Q. (1980). Information in speech: Observations on the perception of [s]-stop clusters. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 536–563.

Baskent, D., & Shannon, R. V. (2003). Speech recognition under conditions of frequency-place compression and expansion. *Journal of the Acoustical Society of America, 113*(Pt. 1), 2064–2076.

Belcher, E., & Hatlestad, S. (1983). Formant frequencies, bandwidths, and Qs in helium speech. *Journal of the Acoustical Society of America, 74,* 428–432.

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America, 114,* 1600–1610.

Best, C. T., Studdert-Kennedy, M., Manuel, S., & Rubin-Spitz, J. (1989). Discovering phonetic coherence in acoustic patterns. *Perception and Psychophysics, 45,* 237–250.

Boersma, P., & Weenink, D. (2000). Praat: Doing phonetics by computer (Version 3.4) [Computer software]. Retrieved October 20, 2000, from the Institute of Phonetic Sciences, University of Amsterdam Web site: www.praat.org

Carhart, R., Tillman, T., & Johnson, K. (1967). Release from masking for speech through interaural time delay. *Journal of the Acoustical Society of America, 42,* 124–138.

Carlyon, R. P., van Wieringen, A., Long, C. J., Deeks, J. M., & Wouters, J. (2002). Temporal pitch mechanisms in acoustic and electric hearing. *Journal of the Acoustical Society of America, 112,* 621–633.

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

Clark, G. M. (2002). Learning to understand speech with the cochlear implant. In M. Fahle & T. Poggio (Eds.), *Perceptual learning* (pp. 147–160). Cambridge, MA: MIT Press.

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

Connine, C. M., & Clifton, C., Jr. (1987). Interactive use of lexical information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 291–299.

Darwin, C. J. (n.d.). Praat scripts for producing Shannon AM speech [Computer software]. Retrieved October 20, 2000, from http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/Praatscripts

Davis, M. H. (2003). Connectionist modeling of lexical segmentation and vocabulary acquisition. In P. Quinlan (Ed.), *Connectionist models of development: Developmental processes in real and artificial neural networks* (pp. 151–187). Hove, United Kingdom: Psychology Press.

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *Journal of Neuroscience, 23,* 3423–3431.

Dorman, M. F., Hannley, M. T., Dankowski, K., Smith, L., & McCandless, G. (1989). Word recognition by 50 patients fitted with the Symbion multichannel cochlear implant. *Ear and Hearing, 10,* 44–49.

Dorman, M. F., Loizou, P. C., & Rainey, D. (1997). Simulating the effect of cochlear-implant electrode insertion depth on speech understanding. *Journal of the Acoustical Society of America, 102*(Pt. 1), 2993–2996.

Dubno, J., Ahlstrom, J. B., & Horwitz, A. R. (2002). Spectral contributions to the benefit from spatial separation of speech and noise. *Journal of Speech, Language and Hearing Research, 45,* 1297–1310.

Dupoux, E., & Green, K. (1997). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 914–927.

Fahle, M., & Poggio, T. (2002). *Perceptual learning.* Cambridge, MA: MIT Press.

Faulkner, A., Rosen, S., & Smith, C. (2000). Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: Implications for cochlear implants. *Journal of the Acoustical Society of America, 108,* 1877–1887.

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003, October 9). Consolidation during sleep of perceptual learning of spoken language. *Nature, 425,* 614–616.

Forster, K. L., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavioral Research Methods Instruments and Computers, 35,* 116–124.

Foulke, E., & Sticht, T. (1969). Review of research on the intelligibility and comprehension of accelerated speech. *Psychological Bulletin, 72,* 50–62.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Science, 3,* 128–135.

Friederici, A. D., Meyer, M., & von Cramon, D. Y. (2000). Auditory language comprehension: An event-related fMRI study on the processing of syntactic and lexical information. *Brain and Language, 74,* 289–300.

Fu, Q. J., & Galvin, J. J., III. (2003). The effects of short-term training for spectrally mismatched noise-band speech. *Journal of the Acoustical Society of America, 113,* 1065–1072.

Fu, Q. J., Shannon, R. V., & Galvin, J. J., III. (2002). Perceptual learning following changes in the frequency-to-electrode assignment with the Nucleus-22 cochlear implant. *Journal of the Acoustical Society of America, 112,* 1664–1674.

Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition, 89,* 105–132.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12,* 613–656.

Gaskell, M. G., & Marslen-Wilson, W. D. (2001). Lexical ambiguity resolution and spoken word recognition: Bridging the gap. *Journal of Memory and Language, 44,* 325–349.

Gathercole, S. E., Willis, C. S., Baddeley, A. D., & Emslie, H. (1994). The Children's Test of Nonword Repetition: A test of phonological working memory. *Memory, 2,* 103–127.

Giraud, A. L., Kell, C., Thierfelder, C., Sterzer, P., Russ, M. O., Preibisch, C., et al. (2004). Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. *Cerebral Cortex, 14,* 247–255.

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

Goldinger, S. D., Kleider, H. M., & Shelley, E. (1999). The marriage of perception and memory: Creating two-way illusions with words and voices. *Memory & Cognition, 27,* 328–338.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology, 49,* 585–612.

Gollin, E. S. (1960). Developmental studies of visual recognition of incomplete objects. *Perception and Motor Skills, 11,* 289–298.

Greenwood, D. D. (1990). A cochlear frequency-position function for several species—29 years later. *Journal of the Acoustical Society of America, 87,* 2592–2605.

Hervais-Adelman, A., Carlyon, R. P., Davis, M. H., & Johnsrude, I. S. (2004). How do cochlear-implant users learn to understand speech? Results from a study using noise-vocoded single words. In *British Society of Audiology Short Papers Meeting on Experimental Studies of Hearing and Deafness* (pp. 147–149). London: University College London.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron, 36,* 791–804.

Jacoby, L. L., Allan, L. G., Collins, J. C., & Larwill, L. K. (1988). Memory influences subjective experience: Noise judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 240–247.

Julesz, B. (1971). *Foundations of the cyclopean perception.* Chicago: University of Chicago Press.

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

Loizou, P. C., Dorman, M., & Tu, Z. (1999). On the number of channels needed to understand speech. *Journal of the Acoustical Society of America, 106*(Pt. 1), 2097–2103.

Luce, P., & Pisoni, D. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19,* 1–36.

Marslen-Wilson, W. (1985). Speech shadowing and speech comprehension. *Speech Communication, 4,* 55–73.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419–457.

Mehler, J., Sebastian-Galles, N., Altmann, G., Dupoux, E., Christophe, A., & Pallier, C. (1993). Understanding compressed sentences: The role of rhythm and meaning. *Annals of the New York Academy of Sciences, 682,* 272–282.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology, 41,* 329–335.

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 39–74). Hillsdale, NJ: Erlbaum.

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

Moore, B. C. J., & Carlyon, R. P. (in press). Perception of pitch by people with cochlear hearing loss and by cochlear implant users. In C. J. Plack, A. J. Oxenham, R. R. Fay, & A. N. Popper (Eds.), *Springer handbook of auditory research.* Berlin: Springer-Verlag.

Morrow, C. (1971). Speech in deep-submergence atmospheres. *Journal of the Acoustical Society of America, 50,* 715–728.

Norris, D. (1993). Bottom-up connectionist models of "interaction." In G. Altmann & R. Shillcock (Eds.), *Cognitive models of language processes: Second Sperlonga Meeting* (pp. 211–234). Hove, United Kingdom: Erlbaum.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52,* 189–234.

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

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

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

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception and 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.

Page, M. (2000). Connectionist modeling in psychology: A localist manifesto. *Behavioural and Brain Sciences, 23,* 443–512.

Pallier, C., Sebastian-Galles, N., Dupoux, E., Christophe, A., & Mehler, J. (1998). Perceptual adjustment to time-compressed speech: A cross-linguistic study. *Memory & Cognition, 26,* 844–851.

Peelle, J., & Wingfield, A. (2003, November). *Adaptation to time-compressed speech by young and older listeners.* Paper presented at the Annual Meeting of the Psychonomic Society, Vancouver, British Columbia, Canada.

Pelizzone, M., Cosendai, G., & Tinembart, J. (1999). Within-patient longitudinal speech reception measures with continuous interleaved sampling processors for ineraid implanted subjects. *Ear and Hearing, 20,* 228–237.

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

Pollatsek, A., & Well, A. D. (1995). On the use of counterbalanced designs in cognitive research: A suggestion for a better and more powerful analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 785–794.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences, 22,* 341–423.

Rauschecker, J. P., & Shannon, R. V. (2002, February 8). Sending sound to the brain. *Science, 295,* 1025–1029.

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

Remez, R. E., Rubin, P. E., Nygaard, L. C., & Howell, W. A. (1987). Perceptual normalization of vowels produced by sinusoidal voices. *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.

Rodd, J., Davis, M. H., & Johnsrude, I. S. (in press). Neural systems for sentence comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex.*

Rosen, S., Faulkner, A., & Wilkinson, L. (1999). Adaptation by normal listeners to upward spectral shifts: Implications for cochlear implants. *Journal of the Acoustical Society of America, 106,* 3629–3636.

Saberi, K., & Perrott, D. R. (1999, April 29). Cognitive restoration of reversed speech. *Nature, 398,* 760.

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.

Schwab, E., Nusbaum, H., & Pisoni, D. B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors, 27,* 395–408.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain, 123,* 2400–2406.

Sebastian-Galles, N., Dupoux, E., Costa, A., & Mehler, J. (2000). Adaptation to time-compressed speech: Phonological determinants. *Perception and Psychophysics, 62,* 834–842.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995, October 13). Speech recognition with primarily temporal cues. *Science, 270,* 303–304.

Shannon, R. V., Zeng, F. G., & Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues. *Journal of the Acoustical Society of America, 104,* 2467–2476.

Skinner, M. (2003). Optimizing cochlear implant speech performance. *Annals of Otology, Rhinology and Laryngology, 191*(Suppl.), 4–13.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002, March 7). Chimaeric sounds reveal dichotomies in auditory perception. *Nature, 416,* 87–90.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 1074–1095.

Summerfield, Q., Haggard, M., Foster, J., & Gray, S. (1984). Perceiving vowels from uniform spectra: Phonetic exploration of an auditory aftereffect. *Perception and Psychophysics, 35,* 203–213.

Toothaker, L. E. (1991). *Multiple comparisons for researchers.* Newbury Park, CA: Sage.

Tyler, R. S., & Summerfield, A. Q. (1996). Cochlear implantation: Relationships with research on auditory deprivation and acclimatization. *Ear and Hearing, 17*(3, Suppl.), 38S–50S.

Voor, J., & Miller, J. (1965). The effect of practice on the comprehension of worded speech. *Speech Monographs, 32,* 452–455.

Warrington, E., & Weiskrantz, L. (1968, March 9). New method of testing long-term retention with special reference to amnesic patients. *Nature, 217,* 972–974.

Watkins, A. (1991). Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America, 90,* 2942–2955.

Weill, S. A. (2001). *Foreign accented speech: Adaptation and generalization.* Unpublished master's thesis, Ohio State University.

Wilson, B. S., Finley, C. C., Lawson, D. T., Wolford, R. D., Eddington, D. K., & Rabinowitz, W. M. (1991, July 18). Better speech recognition with cochlear implants. *Nature, 352,* 236–238.

Received March 8, 2004
Revision received December 22, 2004
Accepted January 7, 2005 ■