Attention, Perception, & Psychophysics
2009, 71 (6), 1360-1374
doi:10.3758/APP.71.6.1360

Effects of intelligibility on working memory demand for speech perception

Alexander L. Francis
Purdue University, West Lafayette, Indiana

and

Howard C. Nusbaum
University of Chicago, Chicago, Illinois

Understanding low-intelligibility speech is effortful. In three experiments, we examined the effects of intelligibility on working memory (WM) demands imposed by perception of synthetic speech. In all three experiments, a primary speeded word recognition task was paired with a secondary WM-load task designed to vary the availability of WM capacity during speech perception. Speech intelligibility was varied either by training listeners to use available acoustic cues in a more diagnostic manner (as in Experiment 1) or by providing listeners with more informative acoustic cues (i.e., better speech quality, as in Experiments 2 and 3). In the first experiment, training significantly improved intelligibility and recognition speed; increasing WM load significantly slowed recognition. A significant interaction between training and load indicated that the benefit of training on recognition speed was observed only under low memory load. In subsequent experiments, listeners received no training; intelligibility was manipulated by changing synthesizers. Improving intelligibility without training improved recognition accuracy, and increasing memory load still decreased it, but more intelligible speech did not produce more efficient use of available WM capacity. This suggests that perceptual learning modifies the way available capacity is used, perhaps by increasing the use of more phonetically informative features and/or by decreasing use of less informative ones.

Under normal circumstances, it seems that we recognize spoken words directly and with so little apparent effort that speech perception has sometimes been considered automatic (Fodor, 1983). Under more challenging circumstances, such as when listening to a talker with an unfamiliar foreign accent, hearing impairment, or dysarthria, or when listening to speech produced by a computer speech synthesizer, speech perception is clearly more demanding; but, subjectively, perception becomes easier with experience. Some support for this subjective perception has been found in studies that showed that training or experience with a talker can significantly improve the intelligibility of dysarthric (Hustad & Cahill, 2003; Liss, Spitzer, Caviness, & Adler, 2002; Spitzer, Liss, Caviness, & Adler, 2000; Tjaden & Liss, 1995), hearing-impaired (Boothroyd, 1985; McGarr, 1983), and foreign-accented (Chaiklin, 1955; Gass & Varonis, 1984) speech, as well as speech produced by a computer text-to-speech (TTS) system (Greenspan, Nusbaum, & Pisoni, 1988; Hustad, Kent, & Beukelman, 1998; Reynolds, Isaacs-Duvall, & Haddox, 2002; Reynolds, Isaacs-Duvall, Sheward, & Rotter, 2000; Rounsefell, Zucker, & Roberts, 1993; Schwab, Nusbaum, & Pisoni, 1985).

One explanation of this perceptual learning is that listeners learn to shift attention to more phonetically informative acoustic cues (Francis, Baldwin, & Nusbaum, 2000), in effect improving the quality of the information being processed by the listener without changing the operation of the mechanism processing that information. It is important to note that this increase is still data-limited by the speech itself; selecting only relevant cues cannot increase the quality of those cues and cannot increase the number of cues or the covariation among them. In essence, this would mean that training allows the listener to more closely approximate the theoretic data limit of a particular synthesizer but cannot raise the asymptote of that limit.

Alternatively, training could improve the way listeners *apply* already-attended acoustic cues to the problem of word recognition, in effect increasing the effectiveness of processing available cues. For example, this could mean that listeners use phonetic context differently for synthetic speech, substantially changing the interpretation of cues. Cues that are initially misleading (due to experience with natural speech) may need to be remapped in order to be interpreted as cuing the intended phonetic category. In terms of Norman and Bobrow's (1975) classification of limitations in information processing, training may thus change data limitations in the first case while changing resource demands in the second. Assuming that the pro-

A. L. Francis, francisa@purdue.edu

© 2009 The Psychonomic Society, Inc.

cessing capacity for speech perception is limited, these two possibilities make different predictions for how training might influence the capacity demands of recognizing synthetic speech.

Recognizing synthetic speech is effortful, possibly because of its impoverished and misleading acoustic cue structure (Nusbaum & Pisoni, 1985). The natural speech signal contains multiple acoustic cues, many of which may indicate the presence of more than one linguistic category. In synthetic speech, this ambiguity may be increased because (1) fewer distinct cues are present (in speech produced by rule) than in natural speech, so the relationships among cues may be uninformative or even misleading, as compared with those in natural speech, and (2) the same patterns of cues can appear in a wider range of contexts (in both rule-based and concatenative synthesis). This acoustic-phonetic ambiguity increases the number of possible phonetic categories that may be activated in working memory (WM), which in turn introduces a greater demand on WM capacity (Cowan, 1988, 1997; Nusbaum & Magnuson, 1997; Nusbaum & Schwab, 1986). For example, in natural speech, a particular cue (e.g., a formant transition with a specific starting frequency, frequency range, and duration) may have a few different possible phonetic interpretations. The alternatives have to be considered and evaluated given other information about speaking rate, talker characteristics, and phonetic and lexical context in order to determine the correct phonetic interpretation. To evaluate the alternatives, they must be represented and tested, thus presenting certain demands on capacity (e.g., WM). The greater the number of alternatives, the more capacity is required (see also Cowan, 1997). For a particular synthetic speech cue in the context of a specific segment, the set of possible phonetic interpretations should be larger than that for natural speech (because acoustic cues in speech generated by rule are generally more acoustically simple and schematic than those in natural speech); therefore, the demands on capacity should be greater.[1]

Moreover, because speech synthesizers are engineered, there can be human engineering errors that result in acoustic properties that are inappropriate for the phonemes they are intended to cue; these properties can be irrelevant or even misleading. In such circumstances, listeners may need to learn to inhibit their attention to certain cues in specific contexts (for evidence of the need for such inhibition in learning a foreign phonetic contrast, see Iverson, Hazan, & Bannister, 2005; Iverson et al., 2003). By withdrawing attention from particular acoustic features that are relevant in natural speech but may be irrelevant or misleading in the speech of a particular TTS system, listeners can reduce the number of phonetic interpretations for a particular set of cues, because the misleading cue is no longer used to generate possible (false) interpretations. Withdrawing attention also allows listeners to free up limited resources for additional processing of other features, because the attention once devoted to the misleading cue can now be redirected and, therefore, be used more productively.

If learning serves to reduce data limitations, listeners should be better able to identify acoustic features that cue appropriate (intended) phonetic categories. Recognition accuracy and response time (RT) should improve, but training should not affect the number of categories that those cues can be interpreted as signaling (i.e., should not reduce the number of alternative phonetic interpretations that are also indicated by the better-identified cues) and, thus, should not alter the capacity demands for recognizing that speech. On the other hand, if training changes the mapping from acoustic cues to phonetic interpretations for a particular TTS system, cues that were not initially perceived as being phonetically diagnostic (because they evoked an unwieldy number of possible phonetic interpretations) may become clearer indicators of phonetic identity; that is, for a given cue, fewer potential phonetic categories may be generated. As the number of possible phonetic interpretations is reduced by the increased validity of the cuing information, recognition accuracy should improve, and capacity demands should drop.

If experience with a particular talker does cause listeners to constrain the set of possible phonetic interpretations of particular acoustic cues better, we would expect to see an interaction between training-related increases in intelligibility and the availability of WM resources. For example, Conway, Cowan, and Bunting (2001) demonstrated that increased WM capacity results in increased selective attention and inhibition of distracting information (see also Engle, 2002; Lavie, 2000; Lavie, Hirst, de Fockert, & Viding, 2004).[2] Training that successfully decreases the number of possible phonetic interpretations of particular acoustic cues should make the deployment of any available resources more efficient, improving accuracy and/or reducing RTs, and should do so to an even greater extent when additional WM capacity is available. Thus, Experiment 1 was designed to investigate the question of whether the WM demands for recognizing synthetic speech would change as a result of training listeners to understand that speech better.

EXPERIMENT 1

Method

Participants. Sixty-one right-handed University of Chicago students and neighborhood residents (24 men and 37 women), ranging in age from 14 to 39 years, all reporting no history of speaking or hearing disorders, were paid to participate in this experiment. All reported English as their native language; 1 participant had prior experience with synthetic speech. Data from 11 participants were excluded from the analyses because those participants failed to complete all of the sessions of the experiment. Twenty-five participants were randomly assigned to the training group, and the remaining 25 were randomly assigned to the untrained control group.

Stimuli. The materials comprised five lists of 50 words each, taken from the phonetically balanced (PB) word list (Egan, 1948). This is a set of 1,000 PB monosyllabic English words, organized in 20 lists of 50 words each, so that the distribution of phonemes within the list as a whole approximates that of English (as does that in each of the sublists, but, necessarily, to a weaker degree). The stimuli were each produced at a natural-sounding rate by a Votrax Personal Speech System and CallText speech synthesizer controlled by an IBM-PC/AT microcomputer. The synthetic speech was low-pass filtered at 4.6 kHz and digitized through a 12-bit analog-to-digital converter at 10 kHz. Each list of words was edited into separate isolated-word stimuli using a digital waveform editor with 100-μsec accuracy and was normalized in root-mean square intensity. Sound files were also equated in duration by adding silence to the end of shorter files. For testing, the digitized words were converted in real time to analog form at 10 kHz with a 12-bit digital-to-analog converter. The speech was low-pass filtered at 4.6 kHz and presented at about 76 dB SPL (measured for the calibration word "cane") over Sennheiser HD 430 headphones. One of the five lists was designated the "practice" list; the other four lists were used in testing.

Procedure. For the training group, the experiment consisted of five 1-h sessions conducted on each of 5 consecutive days (Monday through Friday), or on a single day plus 4 consecutive days (Friday, and Monday through Thursday). For the control group, the experiment consisted of two 1-h testing sessions conducted on Monday and Friday of the same week. The control group received no training sessions at all, because it has been shown that this kind of control group performs identically to a group that receives training on the tasks using natural speech instead of training on those using synthetic speech (Schwab et al., 1985). On the first and last days (Days 1 and 5) of the experiment, the training and control groups were given a word monitoring task to assess the capacity demands of recognizing synthetic speech before and after training. On Days 2 through 4, only the participants in the training group were given training with Votrax-generated synthetic speech.

A speeded word monitoring task was administered to all participants (trained and control) on Days 1 and 5, along with a variant of the WM-load paradigm developed by Baddeley and Hitch (1974). The testing session began with a short practice block of 20 trials from one PB word list (same list used for all listeners), in which listeners monitored for a target word in a series of spoken words. No number preload was given in this practice block, which was followed by two test blocks of 50 trials each. Trials were also blocked by digit load. In one block (low-load), listeners received a number preload of two 2-digit numbers per word list; in the other block (high-load), they received a number preload of five 2-digit numbers per word list. The listeners were instructed to do their best to remember the numbers shown while also responding to the listening task as quickly and accurately as possible (equal weight was given to both tasks in the instructions). The order of the preload conditions was counterbalanced across listeners in each group.

On each trial, the participants were first given a list of two-digit numbers to remember (either two or five numbers, depending on the load condition). Numbers were presented one at a time (i.e., "27," then "46") on a display screen. After seeing the list of numbers, participants saw a target word presented on the display screen and listened for that word to appear in a list of spoken words. They were told to press a response key as quickly and accurately as possible every time they heard the target word. Although the response key was to the right of the computer screen, encouraging responses with the dominant (right) hand, the participants were not explicitly instructed to use only the right hand to respond.

In each trial, listeners heard a list of 16 spoken words produced by a TTS system. Each word was separated from the next by at least 250 msec of silent interstimulus interval (ISI) over and above the silence that was added to shorter files to bring all stimulus files to the same duration. After the monitoring task, the computer prompted participants to use the keyboard to type in the numbers they saw on the screen before the start of the word list presentation, in the order they were presented.

Before each trial began, the computer produced a short tone to alert participants to the beginning of the trial. The word "ready" then appeared on the screen and remained there for 3 sec. After the ready signal, the computer presented a series of randomly selected two-digit numbers. These numbers appeared on the screen one at a time for 2 sec each, with an ISI of 1 sec. After the presentation of the digit list, the target word appeared on the screen. The target word remained on the screen throughout presentation of each series of words. At 3 sec after the target word appeared on the screen, the computer presented a series of 16 spoken words chosen from a single PB list. The target was presented four times at random locations throughout the series of words, but they never appeared first or last in a series, and they never appeared in consecutive positions. The computer selected a different target word for each trial in a block. The target word for one trial could be randomly selected by the computer as a filler word for any of the other trials in the block. In addition, the computer could randomly select the same filler item to appear more than once (but in no more than two consecutive positions) in the same trial.

In all, five PB lists were used for the two test days. One of the five lists was used in the practice block on both test days. Of the remaining four lists, two were used on the first day of testing, and the other two on the last day. Thus, three PB lists were used on each test day. The order of list presentation (excluding the list used for practice on both days) was counterbalanced within each test day.

During Days 2, 3, and 4, the participants in the training group were trained on the identification of spoken words that were produced by a Votrax Type 'N Talk TTS system. On each training trial, the listeners heard a single word and responded by typing the word using a computer keyboard. Immediately following this identification response, feedback was provided about the identity of the stimulus. For this feedback, listeners simultaneously heard the word spoken again and saw the word printed on the computer screen in front of them. The listeners were then asked whether they had identified the word correctly. After the listeners responded Y (yes) or N (no) to this question, the next training trial began. This facilitated scoring right after training and allowed us to check the listeners' confidence in the training procedure. These data were not used in analysis. The listeners were not explicitly told whether they had correctly identified the words or whether they had correctly compared their own identification responses with the feedback they received.

The trainees received three blocks of 50 trials on each training day. (Each block consisted of 50 words from a single PB list.) Although different sets of three PB lists were used each day, the order of list presentation was not varied across participants. Over the course of the 3-day training period, the participants heard a total of 450 novel stimuli from nine PB lists. We did not use any of these lists in testing.[3]

Results

Word identification responses were scored as correct if the response correctly matched the target item phonetically, with no missing, permuted, replaced, or added phonemes. For instance, a response to the word "flew" as "flew," "flue," or "flu" would be considered correct. However, a response of "flute" or "few" or "foo" would be considered incorrect. Results showed that trained listeners improved systematically in word identification performance from 50.2% correct on the first day of training to 55.5% correct on the second day of training and 62.5% correct on the last day of training [$F(2,48) = 94.45$, $p <$ .001].

Number recall. Performance on the WM-load secondary task was scored and analyzed to assess any interaction with the primary speeded word recognition task. In scoring number recall performance, participants received credit for each single digit recalled accurately in the correct sequence. For instance, if the sequence "48 25" was recalled as "48 15," three digits out of four would have been recalled correctly ("4," "8," and "5"). However, if the sequence was recalled as "45 28," credit would be given for only two correctly recalled digits ("4" and "2"). A two-digit number recalled correctly but not in the correct sequence relative to the other two-digit numbers was also counted as correct. For instance, if “23 48” was recalled as “48 20,” credit would be given for recalling “48” correctly (two digits). This allowed a more sensitive measure of recall, although the results were similar in overall pattern to whole-number scoring. Permitting digit credit for recall allowed us to credit more memory-encoding strategies than whole number scoring would have. For example, coding “25” as either “twenty-five” or “two-five” would be scored the same under our method. For listeners using the second memory strategy, recall of either the two or the five would be credited, whereas scoring only for whole numbers would not credit it. We found that our participants tended to use a broader mix of strategies than just pronouncing the whole number, and this scoring was more sensitive to their recall. Mean values (by group and condition) for this variable and for all other variables from all three experiments presented here are shown in Table 1.

In order to confirm that trained listeners did in fact remember more numbers in the high-load condition than in the low-load condition (as instructed), a two-way repeated measures ANOVA with the factors test day (pre, post) and load (low, high) was calculated for digit recall performance. Number recall accuracy was higher when there were fewer digits to remember (97.8% correct) than when there were more (77.8% correct), and this difference was significant [$F(1,24) = 102.17$, $p < .001$]. No other main effects or interactions were significant at the $p < .05$ level. Although the proportion of digits recalled was higher in the low- than in the high-load condition, these proportions represent more digits remembered in the high-load (an average of 7.78 per trial) than in the low-load (an average of 3.91 per trial) condition. Thus, listeners clearly maintained a greater secondary WM load in the high-load condition than in the low-load condition.

Word recognition. Word recognition in the target monitoring task was evaluated in terms of hit rate (proportion of correct recognitions of targets) and RTs for hits. RT was measured from the onset of each stimulus presentation in each trial. If an RT for a particular stimulus was less than 150 msec, the response was assigned to the immediately preceding stimulus in the series, since the time required to recognize and respond to a word is unlikely to be faster than 150 msec. The RT for this preceding stimulus was computed as the duration of the stimulus plus the ISI plus the recorded RT of less than 150 msec measured for the following stimulus.

In order to determine whether training was effective, as suggested by the results of the training-day analysis described above, a three-way repeated measures ANOVA of hit rate with the factors of group (trained vs. control), test day (pre, post), and WM load (low, high) was computed. Results showed a significant effect of test (pre = 89.1%, post = 91.5%) [$F(1,48) = 10.26$, $p = .002$] and of load (low = 91.5%, high = 89.0%) [$F(1,48) = 15.27$, $p < .001$], as well as a significant interaction between group and load [$F(1,48) = 11.29$, $p = .002$]. Post hoc (Tukey’s HSD) analysis showed a significant ($p < .05$) difference between performance on the pretest and on the posttest only for the trained group (pre = 88.9%, post = 94.2%) and not for the control group (pre = 89.3%, post = 88.7%), suggesting that training was the major source of listeners’ improved performance on the word recognition task.

In order to further explore the effect of training on the use of available WM capacity, a two-way repeated measures ANOVA was calculated for the hit rate scores of trained listeners only, with the factors of test day (pre, post) and WM load (low, high). The results showed a significant effect of test [$F(1,24) = 26.13$, $p < .001$], with trained listeners improving from 88.9% correct on the pretest to 94.2% correct on the posttest. There was also a significant effect of load [$F(1,24) = 8.53$, $p = .007$], with greater accuracy in the low-load (93.0%) than in the high-load (90.1%) condition. However, there was no interaction between WM load and the effects of training on recognition accuracy [$F(1,24) = 0.69$, n.s.].

A two-way repeated measures ANOVA of trained listeners’ RTs, with the factors of test day (pre, post) and WM load (low, high), showed a significant effect of test day [$F(1,24) = 6.38$, $p = .02$] and load [$F(1,24) = 21.73$, $p < .001$], as well as a significant interaction between the two [$F(1,24) = 4.46$, $p = .045$] (see Figure 1). Examining

Table 1
Data Summary for the Low and High Working Memory Load Conditions in Experiments 1, 2, and 3

		Word Recognition															
		Number Recall				Hit Rate				False Alarms				RT			
		Low		High		Low		High		Low		High		Low		High	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Experiment 1	CPre	.97	.04	.75	.13	.90	.07	.88	.08	2.00	2.02	2.32	1.68	608	86.72	609	76.86
	CPost	.95	.09	.77	.18	.90	.09	.88	.08	2.32	2.10	3.64	2.58	606	87.83	618	86.72
	TPre	.98	.03	.77	.10	.91	.08	.87	.09	1.84	1.65	2.56	2.36	601	84.21	621	91.94
	TPost	.98	.03	.79	.13	.95	.04	.93	.06	2.20	1.63	2.12	1.92	558	92.21	602	82.22
Experiment 2	Low intel	.91	.08	.60	.14	.86	.08	.86	.08	8.59	5.86	10.72	5.54	450	38.81	437	40.41
	High intel	.93	.08	.61	.16	.94	.05	.90	.13	3.97	3.54	6.10	4.71	456	5.30	447	63.30
Experiment 3	Low intel	.98	.03	.75	.13	.91	.07	.87	.11	2.71	2.26	2.95	2.13	595	95.51	615	81.97
	High intel	.97	.04	.78	.15	.97	.04	.91	.07	1.38	1.02	2.00	1.90	519	81.20	534	78.67

Note—RT, response time; CPre, control group pretest; CPost, control group posttest; TPre, training group pretest; TPost, training group posttest; Low intel, low intelligibility talkers; High intel, high intelligibility talkers. Number recall and hit rate are given as proportions of 1. False alarms are given in terms of average number of false alarms per condition (rather than proportion), due to their relative rarity. RTs were recorded for correct responses only.

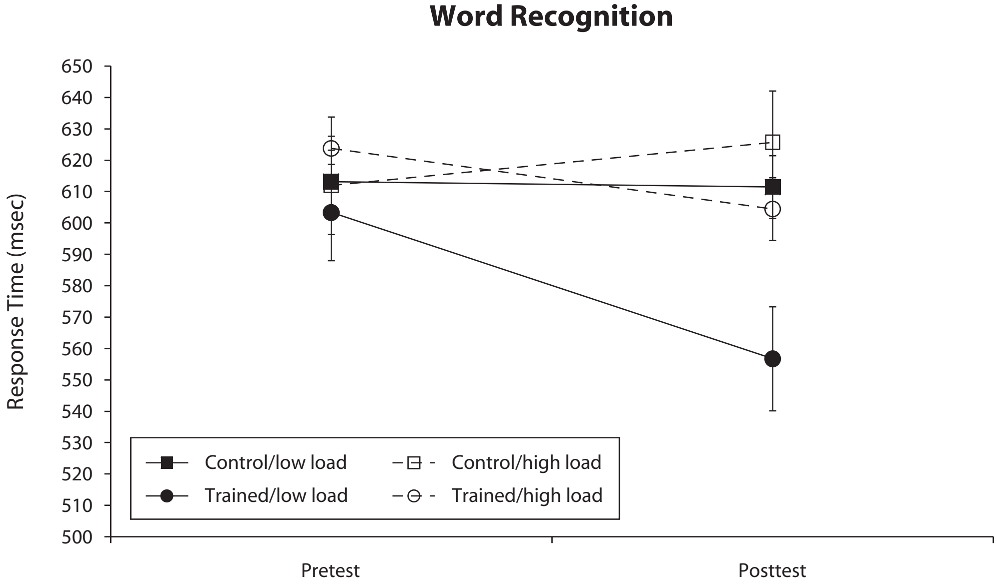

Figure 1. Experiment 1: Response time for correctly recognized target words in each memory load condition, as a function of training on synthetic speech. Bars delineate ±1 standard error of the mean.

the interaction, post hoc (Tukey's HSD, critical p value for reporting significance = .05) analysis indicated that load had no significant effect on RT on the pretest (low, 601 msec; high, 621 msec) but did on the posttest (low, 558 msec; high, 602 msec). More important, training significantly reduced listeners' RTs under low WM load (from 601 msec on the pretest to 558 msec on the posttest), but the change in RT was not significant under high load (from 621 msec on the pretest to 602 msec on the posttest).

Relationship between variables. Correct interpretation of the results presented thus far depends on the assumption that participants did not change their resource allocation strategies in a load- or test-day-dependent manner. Analysis of the correlation between number recall and word recognition accuracy for the trained group showed no correlation at either low or high load on either the pretest or the posttest, with R^2 values ranging from .003 to .12 and p values ranging from .10 to .79. However, there were significant but relatively small correlations between word recognition accuracy and RTs on correct responses at both low and high load on both the pretest and the posttest. In this case, R^2 values ranged from .16 to .45, with p values ranging from .0002 to .05. However, such correlations are not unexpected; Nusbaum and Pisoni (1985) found that words that are recognized more accurately are also recognized more quickly.

Recall of numbers under high WM load at pretest should have reflected individual differences in WM capacity and memory strategy, and recall performance under high WM load at posttest should have reflected the participants' capacity and strategy after training. Because the number of digits to be remembered in the low-load condition was relatively small, the low WM-load recall performance should not have differed much among participants. Thus, comparing the difference between low- and high-load performance from pretest to posttest should reflect any shift in recall strategy that might differ between groups (perhaps as a confound with training) that could affect speech recognition performance.

Regression of pretest and posttest recall performance as independent variables to predict the pretest–posttest RT difference shows that neither measure of memory performance was significantly related to speech recognition speed (i.e., neither coefficient was a significant predictor of the RT difference; for pretest recall, $t = 0.69$; for posttest recall, $t = -0.14$). Indeed, there was no significant difference in recall of the high WM load from pretest to posttest between the control and trained groups [$F(1,48) = 0.03$, n.s.], suggesting that both groups showed the same recall of the high WM load at pretest and posttest.

Even though there was no systematic change in WM recall from pretest to posttest between the groups, and overall there was no reliable relationship between either pretest recall or posttest recall and the effect of training on speed of speech recognition, it is possible that other, unidentified, systematic differences in WM performance within each of the two groups (control, trained) might account for the training effects observed here. To test this, we repeated the ANOVA on word recognition speed using WM recall (at high load) as a covariate. If differences between the groups in use of WM are systematic and related to the training effect on speech recognition, this covariate should mitigate the group difference. To do an ANCOVA to take into account the recall abilities of participants, we simplified the ANOVA on RT. The basic interaction occurs because there was a difference in RT from pretest to posttest for the training group but not for the control group, and this difference was seen only at low load. That is, the effect of interest is in the RT difference scores between the pretest and posttest across the two groups in the low-load condition.

Therefore, if we look only at the low-load condition and compute a score consisting of the difference in RT from pretest to posttest, there should be a significant difference in this score between groups, and there was [$F(1,48) = 7.74$, $p < .008$]. Moreover, the same test at high load was not significant [$F(1,48) = 2.69$, n.s.], which shows the nature of the interaction. Taking the differences in WM into account as a covariate does not eliminate this interaction of training with WM load. Neither individual differences in WM capacity between participants or between groups nor shifts in dual-task performance between participants or groups can account for the interaction between WM load and training effects in the perception of synthetic speech. When the listeners are trained on synthetic speech, at the low WM load, they can use the residual surplus capacity to speed up word recognition even more than they can at the high WM load.

Discussion

When we consider the effects of training on recognition accuracy observed in the speeded word recognition task, the present results generally accord well with the results of previous studies of synthetic speech learning using a word identification task (e.g., Schwab et al., 1985). Training significantly improved the intelligibility of synthetic speech, as measured by recognition accuracy, whereas untrained control participants showed no such improvement. This, taken by itself, is compatible with a variety of auditory theories of speech perception (e.g., Cleary & Pisoni, 2001; Diehl, Lotto, & Holt, 2004; Goldinger, 1998). Furthermore, since there was an effect of WM load on recognition accuracy, these results support the more specific hypothesis that spoken word recognition depends in part on the availability of WM capacity (see also Luce, Feustel, & Pisoni, 1983).

A similar pattern of results was found for RT measurements, a variable that is more sensitive to real-time cognitive demands on processing. Both before and after training, word recognition was slower when there were greater WM demands imposed by the number preload task. Furthermore, training increased the speed of word recognition. Of particular interest, however, is the significant interaction between training-induced changes in intelligibility and WM load. Specifically, in the low-load condition, training resulted in significantly faster RTs, whereas in the high-load condition, training had much less of an effect on RTs. These results suggest that perceptual learning of synthetic speech may change the way available WM capacity is used for the task of speech perception (cf. Baddeley, 2002; Baddeley & Hitch, 1974).

In terms of resource theory (Lavie, 2005; Norman & Bobrow, 1975), one interpretation of the present results is that learning alters the demands of a resource-limited process. WM capacity is limited, but training improves the manner in which it is allocated or used, thereby improving processing when capacity is available. On the other hand, these results are also consistent with a model of learning that is based on changing the quality of the input to a data-limited process by reducing the number of misleading cues that are processed. According to this hypothesis, training improves intelligibility of the synthetic signal by directing listeners' attention away from uninformative acoustic cues and toward more informative ones (Francis, Nusbaum, & Fenn, 2007). In turn, the improved signal becomes easier to process and thus requires a reduced commitment of available WM resources.

Recent research on speech perception in noise and by older people suggests that understanding low-intelligibility speech can be limited by the availability of WM capacity even when the decrease in intelligibility is purely a consequence of limiting the quality of the acoustic input. McCoy et al. (2005), Pichora-Fuller, Schneider, and Daneman (1995), and Rabbitt (1991) have argued that understanding poorly heard speech (at low signal-to-noise ratio or with hearing impairment) requires the commitment of additional processing resources (e.g., WM), and that committing these resources to the task of speech recognition results in a reduction of the WM resources available for subsequent (linguistic, conceptual) processing of the recognized speech. It is possible that the changes in WM demands observed in Experiment 1 were derived directly from changes in intelligibility, rather than from changes in the listeners' ability to process the acoustic cue structure of a specific speech synthesizer. In other words, although it is indisputable that training improved the intelligibility of the synthetic speech used in this experiment, this is not definitive evidence that training in and of itself affected WM demands for understanding this synthesizer. It is possible that the observed changes in WM demand were a direct consequence of the change in intelligibility; thus, they might be observed as a consequence of *any* change in intelligibility, even if it derives from some process other than perceptuo-linguistic experience. In order to test this hypothesis, Experiment 2 was designed to test the interaction of WM demands with intelligibility by testing listeners with two different computer speech synthesizers differing in intelligibility, thus retaining differences in intelligibility but eliminating training effects.

EXPERIMENT 2

TTS systems can differ widely in intelligibility (Hustad et al., 1998; Mirenda & Beukelman, 1990; Nusbaum & Pisoni, 1985); therefore, switching between two different synthesizers provides a way of manipulating intelligibility without changing the listener. In this experiment, participants completed two blocks of testing using a modified version of the speeded target monitoring task combined with the same secondary memory-load task used in Experiment 1. The two blocks were identical in structure but differed in terms of the synthetic talker used to produce the speech the listeners heard.

Method

Participants. Forty-two native students and community members from Purdue University were recruited to participate in this experiment (28 women and 14 men, ranging in age from 18 to 29 years). All of the participants reported speaking English as their native language. Of these, 3 (2 men, 1 woman) were excluded either because they had significant, early experience with one or more languages other than English (1 participant had 2 years' experience living in a non-English environment before the age of 12) or because they did not return to complete the second day of testing (2 participants). All of the participants passed a pure-tone audiometric screening at 25 dB HL at 500 Hz, and at 20 dB HL at 1, 2, and 4 kHz. The participants were paid for their participation under a protocol approved by the committee for human research subjects at Purdue University.

Stimuli. The stimuli comprised four separate test sets and two practice sets of words drawn from the PB word lists (Egan, 1948). Rather than select from the full set, as in Experiment 1, we used subsets to simplify the procedures for stimulus generation and presentation. Each test set comprised 20 trials of 16 tokens each. All of the tokens were drawn from the same randomly selected PB word list for a given set (Lists 15, 19, 9, and 4 for Sets 1, 2, 3, and 4, respectively). Each trial contained 16 words, including four instances of a randomly selected target. The occurrences of each target were randomly selected from among the 16 possible entries in the list, with the constraints that at least one distractor must be presented between instances of the target and that the target could not appear in the first or last position in the list. The other 12 positions within each trial list were filled with randomly selected words from the PB list. Multiple instances of a particular distractor within a particular trial list were allowed, as were instances of a target from 1 trial appearing as a distractor in another trial. Two practice lists were also generated, each containing 5 trials of words drawn from two additional randomly selected PB lists (7 and 14), and were constructed according to the same principles as the test lists.

Each set of words (four test and two practice) was produced by each of two different computer speech synthesizers. One synthesizer, rsynth (Ing-Simmons, 1994), is a publicly available, rule-based, cascade/parallel formant synthesizer based on the Klatt engine (Klatt & Klatt, 1990); the other is an inexpensive concatenative synthesizer (Cepstral, 2006). These talkers were selected because they are roughly comparable in overall intelligibility to Votrax and DECtalk[4] but are more easily accessible than the older synthesizers. Stimuli produced by rsynth had a mean duration of 339 msec (*SD* = 76 msec), whereas those produced by Cepstral David had a mean duration of 472 msec (*SD* = 82 msec).

Procedure. The experiment consisted of two sessions on one day and two sessions on another. Each session was comparable to the pretest session of Experiment 1. On the first day, listeners completed a short hearing test and a linguistic background questionnaire before being seated individually in study carrels in a quiet room. Each carrel contained a computer (Dell Dimension running Windows XP) using E-Prime 1.2.1 (Schneider, Eschman, & Zuccolotto, 2002) and a Cedrus RB-620 response box for presenting stimuli and recording responses. The stimuli were presented through Sennheiser HD 25 headphones.

On the first day of testing, the participants were randomly assigned to hear either rsynth stimuli or Cepstral David stimuli; they heard the other talker on the second day. On each day, listeners first completed a short practice session identical to the test session but without a memory preload task and consisting of only five trials (see above), to become familiar with the talker and the target monitoring task. There were two practice lists, one for each day, generated with the same synthesizer used in testing. Practice sessions were not scored.

After completing the practice session, listeners completed 2 blocks of trials. One block was completed under high load (remembering five 2-digit numbers) and the other under low load (two 2-digit numbers). Load block order was counterbalanced across participants, and the instructions emphasized giving equal weight to the performance of the two tasks (memory and listening). In each block, the listeners completed 20 speeded target monitoring trials, each drawn randomly from a single set (see Stimuli, above) and produced by a single talker. Each trial consisted of the presentation of a total of 16 words, four targets and 12 distractors, as described above. Although the order of trials was randomly selected for each participant, the order of stimuli within a given trial was the same for all participants. On each trial, the participants first saw a signal to get ready, followed after 500 msec by the numbers for the memory-preload task. These numbers were presented visually, one at a time (2 digits per number). Each number remained on the screen for 2 sec, with a 1-sec ISI. At 1 sec after the presentation of the last number, the target word that the listeners were asked to monitor for was displayed visually, and 500 msec after that the first word was presented auditorily. As in Experiment 1, the words were presented with a variable ISI (never less than 250 msec) to maintain a consistent SOA (940 msec), meaning that each monitoring block was the same duration both within and across synthesizers, in order to maintain a consistent span for the memory-load task.

On the second day of testing, the listeners completed another two blocks of testing identical to those described for the first day, except that each block used a different stimulus set and the other practice set was used. Thus, four sets of stimuli were used: one for each block (low and high memory load) for each talker/day (rsynth, Cepstral David). The selection of sets for each talker and block was counterbalanced across listeners. In this way, the listeners never heard the same word used as a target in any trial, regardless of talker or day, but did hear target words from one trial used as a distractor in another trial within the same load and talker condition.

Results

Number recall. Performance on the WM-load secondary task was scored and analyzed as in Experiment 1. The results showed a significant effect of load [$F(1,38) = 338.95, p < .001$]. Mean recall accuracy was higher under low load (92.1%, nearly four digits) than when there were more visually presented numbers to recall (60.5%, just over six digits). Results showed no effect of talker on the WM task [$F(1,38) = 2.28, p = .14$], meaning that recall accuracy did not differ significantly between the rsynth (75.5%) and Cepstral David (77.0%) speech conditions, and there was no significant interaction between type of speech and memory load [$F(1,38) = 0.54, p = .47$]. This shows that, as in Experiment 1, speech intelligibility did not affect how memory was allocated to the secondary number recall task.

Word recognition. Accuracy and RT were computed and analyzed as in Experiment 1. Accuracy showed a significant effect of talker [$F(1,38) = 21.38, p < .001$], with listeners responding more accurately to Cepstral David (92.0%) than to rsynth (89.2%), but no significant effect of load (low, 90.1%; high, 87.8%) [$F(1,38) = 2.51, p = .12$]. There was, however, a marginally significant interaction between talker and load [$F(1,38) = 3.97, p = .053$]. This interaction was highly significant using arcsine-transformed data[5] [$F(1,16) = 7.43, p = .009$], and post hoc (Tukey's HSD) analysis of the arcsine-transformed results showed a significant effect of load ($p < .05$) only in the high-intelligibility talker (high, 89.7%; low, 94.4%). However, the high-intelligibility talker was perceived significantly more accurately ($p < .05$) than the rsynth (low, 85.8%; high, 86.0%) in both the low- and the high-load conditions. This pattern of interaction in hit rate is the same as that observed in RT in Experiment 1, although no such interaction was observed in accuracy in Experiment 1, and no corresponding interaction in RT was observed in Experiment 2.

With respect to RT, there was a small but significant effect of load [$F(1,38) = 7.53, p < .01$], but the direction of the effect was unexpected, with listeners responding correctly somewhat more slowly under low load (453 msec) than under high load (442 msec). There was no effect of talker [$F(1,38) = 2.14$, $p = .15$] and no interaction [$F(1,38) = 0.15$, $p = .70$]. It is not clear why the listeners were able to respond more quickly under high load than under low load; given the observed lack of a significant effect of load in the accuracy data, this could indicate a speed–accuracy trade-off, with listeners responding more quickly at the expense of response accuracy, at least in the low-load condition. However, hit rate did not correlate with RT in either the low- or high-load conditions in Experiment 2.

Effects of talker order. One interesting difference between this experiment and the first is that, in this case, half of the participants heard the low-intelligibility voice first, whereas the other half heard the high-intelligibility voice first. Because the first experiment was a training study, all of the participants necessarily completed the low-intelligibility test session before the high-intelligibility one. Thus, in order to obtain the closest comparison with Experiment 1, we reanalyzed number recall and word recognition accuracy and RT for only those participants who also heard the low-intelligibility voice (rsynth) first, followed by Cepstral David ($n = 17$).

Analysis of these listeners' number recall showed a significant effect of talker [$F(1,16) = 4.78$, $p = .044$] and of load [$F(1,16) = 112.87$, $p < .001$], but no interaction [$F(1,16) = 0.23$, $p = .64$]. Accuracy (hit rate) data showed no significant effects of load [$F(1,16) = 2.06$, $p = .17$] or talker [$F(1,16) = 2.46$, $p = .14$], nor any interaction [$F(1,16) = 3.31$, $p = .09$], although the participants trended toward recognizing more words in the low-load (90.5%) than in the high-load (86.5%) condition and toward recognizing more words produced by Cepstral David (90.6%) than those by rsynth (86.3%). Arcsine-transformed hit rates were analyzed similarly, showing a significant effect of talker [$F(1,16) = 7.64$, $p = .01$], but no significant effect of load [$F(1,16) = 3.23$, $p = .09$]. There was, however, a significant interaction between talker and load [$F(1,16) = 5.87$, $p = .03$]. In this case, post hoc (Tukey's HSD) analysis indicated that this interaction reflected a significant ($p < .05$) difference between the talkers *only* under low memory load (rsynth/low-intelligibility, 86.0%; Cepstral David/high-intelligibility, 94.9%) and not under high load, unlike the overall group data. Moreover, load affected only the Cepstral David (high-intelligibility) voice, so that the change from low to high load decreased hit rate for the David voice from 94.9% to 86.3%, whereas the same difference for rsynth (from 86.0% to 86.6%) was not significant. RT showed no significant effects for load [$F(1,16) = 1.01$, $p = .33$] or talker [$F(1,16) = 0.08$, $p = .79$] and no interaction [$F(1,16) = 0.03$, $p = .87$], and none of these effects became significant using log-transformed data. Thus, these participants showed exactly the same pattern in their hit rate data as the participants in Experiment 1 showed in their RT results, although the RT results in the present case do not show a similar pattern. That is, in both the Experiment 1 RT data and the present accuracy data (for listeners exposed to rsynth followed by Cepstral David), load showed an effect only in the high-intelligibility context (after training in Experiment 1, and with the Cepstral David talker here); but changing intelligibility, whether by training or by changing talkers, showed a significant effect only under low load.

Interestingly, this pattern was *not* observed in the participants who heard the high-intelligibility talker first ($n = 22$). A two-way repeated measures ANOVA of these listeners' hit rate showed a significant effect of talker [$F(1,21) = 57.64$, $p < .001$], but not of load [$F(1,21) = 0.47$, $p = .50$], and no significant interaction [$F(1,21) = 0.81$, $p = .38$]. Hit rate was higher for the Cepstral David voice (93.1%) than for rsynth (85.6%). A similar analysis of their RTs showed a significant effect of talker [$F(1,21) = 5.54$, $p = .03$] and load [$F(1,21) = 8.01$, $p = .01$], but no significant interaction [$F(1,21) = 0.32$, $p = .58$]. Correct responses were faster for rsynth heard in the second session (439 msec) than for Cepstral David heard in the first session (455 msec). These listeners were also faster under high (440 msec) than under low (454 msec) load. Thus, the group results described above are really only indicative of the performance of the participants who heard the low-intelligibility talker first. This suggests that the observed pattern of interaction between intelligibility and load may result in large part from order (learning) effects in this experiment, as well as from that in the first one.

Discussion

The observed increase in RT with increased intelligibility is unexpected, given that previous research has shown that listeners are quicker to recognize more intelligible speech (Nusbaum & Pisoni, 1985). In this case, there are two possible explanations for this pattern. (1) This increase may be due to the much greater average duration of the Cepstral David stimuli, as compared with those produced by rsynth, so that listeners take longer to respond to the David voice simply because the stimuli are, themselves, longer. (2) It is also possible that this pattern results from differences in synthesis methods: The Cepstral David TTS is a concatenative synthesizer, whereas rsynth is a Klatt-style cascade/parallel formant synthesizer. This means that the Cepstral stimuli contain many more of the acoustic properties found in the natural voice from which they are derived, whereas those stimuli produced by rsynth contain only those acoustic patterns specifically introduced by the system's designer. Thus, the greater acoustic richness of the Cepstral stimuli may also have increased the number of acoustic cues listeners could process, leading not only to improved performance but also to greater processing time, as compared with the more impoverished formant-synthesized rsynth stimuli.

Differences between synthesizers might also explain the observation that, in Experiment 2, manipulation of WM load only *seemed* to affect performance on the Cepstral David stimuli, and that differences between talkers are only seen under low load. It is possible that the acoustic richness of the Cepstral David voice may have increased the listeners' susceptibility to load effects, even though the increased availability of a wider variety of acoustic cues contributed to this talker's superior intelligibility overall because the listeners had distributed their attention to more acoustic cues in the Cepstral David stimuli than to those in the acoustically simpler rsynth stimuli. Similarly, differences between talkers may have been more apparent under low load because, under high load, listeners had insufficient capacity available to make maximal use of the greater variety of acoustic cues available in the Cepstral David stimuli. Only as surplus capacity became available in the low-load condition could the listeners benefit from distributing more attention to more available cues.

These differences aside, the results of Experiment 2 are consistent with the hypothesis that the experience of hearing a poorly intelligible voice predisposes listeners to attend flexibly to the acoustic properties of speech (Francis et al., 2007). According to this argument, listeners exposed to less intelligible speech recognize the ineffectiveness of the cues they typically attend to and shift their attention to cues that may be more diagnostic for a particular talker. It is possible that there are two components to this process: an initial release of attention to typical cues and a subsequent reorientation toward more effective ones. Listeners who were first exposed to a low-intelligibility talker may have released attention from a greater number and/or variety of typical cues (since lower intelligibility indicates that more of their typical cues have been proven ineffective), whereas the listeners who were first exposed to a highly intelligible, natural-sounding talker may have released fewer cues. In other words, although listeners in Experiment 2 were not specifically trained to better understand a particular voice, it is possible that the sequential order of the test conditions may have encouraged behavior comparable to some of the processes involved in learning.

EXPERIMENT 3

The third experiment was carried out as a replication of Experiment 2 but used a between-participants design in order to eliminate the possibility of test-order effects' influencing the effect of intelligibility on load. To further explore the generality of the load effect, and to eliminate the possibility that listeners' performance might have been influenced by attributes of the specific synthesizers in Experiment 2, two different synthesizers were used: the Votrax Type 'N Talk system used in Experiment 1 and a more intelligible synthesizer, CallText 5000. These two synthesizers closely match the pretest and posttest intelligibility scores in Experiment 1. By comparing the perception of Votrax speech with the perception of CallText speech, we were able to investigate the effects of changing intelligibility to about the same degree as we had observed them in Experiments 1 and 2, but without training and without even repeating the experience of listening to synthetic speech. If the load effects observed in Experiments 1 and 2 were merely the result of changes in intelligibility—and not of experience or learning—they should be observed in Experiment 3 as well.

Method

Participants. Forty-seven University of Chicago students and community residents (26 men and 21 women, ranging in age from 17 to 33 years) were paid to participate in Experiment 3. All of the participants reported speaking English as their native language, and none reported speech or hearing disorders (no formal hearing screening was conducted). The data from 5 participants were excluded from the analyses. Of these 5, 3 participants were excluded because of equipment failure, 1 because her native language was not English, and 1 because he was unable to perform the task.

Stimuli. The stimuli comprised the same sets of words produced by the Votrax Personal Speech System as in Experiment 1 and an identical set produced by the CallText speech synthesizer controlled by an IBM-PC/AT microcomputer. Mean stimulus durations for PB words produced by Votrax and CallText were 446 msec (*SD* = 73 msec) and 295 msec (*SD* = 79 msec), respectively.

Procedure. The experiment consisted of a single 1-h session. The participants were assigned to one of two experimental conditions. The first group listened to speech produced by the less intelligible Votrax system. The second group listened to the more intelligible speech produced by the CallText system. Other than the choice of talker, the test session was identical to the posttest session of Experiment 1.

Results

Number recall. Performance on the WM-load secondary task was scored and analyzed as in Experiment 1 in order to assess any interaction with the primary speeded word recognition task. Results of a two-way mixed factorial ANOVA with one between-participants factor (Votrax, CallText) and one within-participants factor (low load, high load) showed a significant effect of load [$F(1,40) = 112.14, p < .001$], showing that mean recall accuracy was higher when there were fewer numbers to recall (97.3%, nearly four digits) than when there were more visually presented numbers to recall (76.6%, nearly eight digits). Results showed no effect of talker [$F(1,40) = 0.15, p = .70$], meaning that recall accuracy did not differ significantly between the Votrax (86.5%) and CallText (87.4%) speech conditions, and there was no significant interaction between type of speech and memory load [$F(1,40) = 1.43, p = .24$]. As in Experiment 1, these results again suggest that the intelligibility of the speech did not affect how listeners allocated WM to the secondary number recall task.

Word recognition. For the word-recognition task, we computed proportion of correct detections (hits) and RTs for correct responses, as was done in Experiment 1. Results of a two-way mixed factorial ANOVA of accuracy ratings showed a significant effect of load [$F(1,40) = 13.81, p = .001$], showing that participants recognized more words in the low-load condition (93.9%) than in the high-load condition (89.5%). There was also a significant effect of talker [$F(1,40) = 5.46, p = .02$]; listeners recognized more words produced by CallText (94.1%) than by Votrax (89.2%). Taken together, these results are compatible with the findings of Pichora-Fuller et al. (1995) and Rabbitt (1991), showing that speech perception accuracy was higher for the more intelligible speech and decreased with increased memory load. However, the interaction between type of speech and WM-load condition was far from significant [$F(1,40) = 0.49, p = .49$].

Results of a two-way mixed-factorial ANOVA of RTs showed a significant effect of load [$F(1,40) = 6.49, p = .01$]. Listeners were significantly faster at recognizing words in the low-load condition (557 msec) than in the high-load condition (575 msec). There was also a significant effect of talker [$F(1,40) = 9.73, p = .003$]. Listeners recognized words faster for CallText (527 msec) than for Votrax (605 msec) speech, perhaps in part because the CallText stimuli were much shorter (295 msec for CallText vs. 446 msec for Votrax). However, unlike in Experiment 1, in which the change in intelligibility was accomplished by training, there was no significant interaction between the level of load and type of speech [$F(1,40) = 0.17, p = .68$], nor even the appearance of a trend in this direction, although both main effects were clearly significant.

Discussion

Listeners were both faster and more accurate under conditions of low load and when listening to more intelligible speech, and there was no interaction between the two factors (load and intelligibility) for either accuracy or RT data. Recognizing poorly intelligible speech requires more time and is accomplished less accurately than recognizing more intelligible speech. In a manner complementary to the results of Pichora-Fuller et al. (1995), Rabbitt (1991), and McCoy et al. (2005), who found that decreasing intelligibility reduced the WM capacity available for other tasks, such as encoding and understanding, the present results support the hypothesis that recognition of speech is sensitive to the availability of WM resources by showing that, when WM resources are limited (in the high-load condition) recognition performance suffers, regardless of intelligibility. However, the lack of an interaction between intelligibility and load suggests that the processing demands incurred by these two factors may be independent. That is, whatever is causing the increased RT and lowered accuracy for poorly intelligible speech does not appear to be affected by the availability of WM resources. This suggests that listeners have prioritized the WM task, perhaps for intrinsic reasons or, perhaps, because, despite all efforts, they perceived it as being more important within the overall context of the experiment. It also suggests a distinction between the specific WM resources that each task draws upon, so that the memory buffer involved in rehearsal for the digit task may be distinct from that used in lexical access (Rochon, Caplan, & Waters, 1990).

GENERAL DISCUSSION

The significant results for all three experiments are summarized in Table 2. Low-intelligibility synthetic speech is not just less well understood than high-intelligibility synthetic speech, it is typically understood more slowly. Furthermore, limiting the availability of WM generally increases RT and decreases accuracy for the recognition of synthetic speech. Exceptions to these observations are difficult to interpret and may result from specific differences in the synthesis methods used for specific voices. More significantly, the two factors do not appear to interact when the change in intelligibility is the result of a difference between talkers, except when this difference mirrors the effects of training by presenting the more intelligible speech after the less intelligible. Crucially, when less intelligible synthetic speech is made more intelligible by training listeners, there is an interaction between intelligibility and the availability of WM. In particular, training causes listeners to become much faster at recognizing synthetic speech under conditions of low memory load, as compared with conditions of higher memory load. This suggests that training specifically involves changing some aspect of the way WM resources are applied to the task of speech recognition.

Perceptual learning of synthetic speech involves learning about the mapping between the acoustic-phonetic properties of the synthetic speech and the categories they are intended to cue, allowing further generalization to novel utterances produced by that synthesizer. This is demonstrated by the improvements in recognition accuracy in the present study, as well as by the improvements in the identification of novel words observed in previous experiments (e.g., Francis et al., 2000; Greenspan et al., 1988; Schwab et al., 1985). Furthermore, our previous research has demonstrated that listeners learn to shift at-

Table 2
Summary of Results for Experiments 1, 2, and 3

		Experiment 2		
	Experiment 1	Overall Group	rsynth First	Experiment 3
Training Results	pre < post	N/A	N/A	N/A
Number Recall PC	high < low**	high < low**	high < low** rsynth < David	high < low**
Speeded Target Monitoring Accuracy	pre < post high < low	rsynth < David*	rsynth < David*	Votrax < CallText high < low
Interaction*		David: high < low low: rsynth < David high: rsynth < David	David: high < low low: rsynth < David	
Speeded Target Monitoring RT	post < pre low < high	high < low		CallText < Votrax low < high
Interaction	post: low < high low: post < pre			

Note—All results for Experiment 1 are for trained listeners only. PC, proportion correct; RT, response time; pre, pretest; post, posttest; low, low load; high, high load; David, Cepstral David text-to-speech (TTS) talker; rsynth, rsynth TTS talker; Votrax, Votrax Type 'N Talk TTS talker; CallText, CallText TTS talker. Only significant comparisons are shown. *Only significant after transformation (arcsine). **These results notwithstanding, more digits were recalled under high load.

tention to acoustic properties of speech that are given less weight in perceptual classification prior to training (Francis et al., 2000) or are not used at all (Francis & Nusbaum, 2002; in the visual modality, see also Goldstone & Schyns, 1994; Livingston, Andrews, & Harnad, 1998), modifying the structure of the mapping between acoustic-phonetic cues and phonetic categories (Francis et al., 2007). Listeners can use comparatively high-level feedback about the intended categorization of an utterance (feedback only indirectly related to the specific acoustic properties that must be learned) to direct attention to the speech signal in order to achieve that categorization. This is consistent with theories that consider perceptual learning in terms of the redistribution of attention among perceptual cues (e.g., Goldstone, 1994; Nosofsky, 1986).

The present results demonstrate that training also produces changes in the way perception of synthetic speech uses WM. After training, listeners used available WM capacity more effectively in recognizing synthetic speech. The more capacity they had available, the greater their improvement in speed of recognition. This pattern of results cannot be accounted for simply in terms of improving the intelligibility of perceiving synthetic speech. Training produces an effect on the use of WM that simply improving intelligibility does not produce.

The effects of perceptual learning on the use of WM and the effects of learning on attention may be closely related through the relationship between attention and WM (e.g., Bundesen, 1990; Cowan, 1988, 1997; Nusbaum & Schwab, 1986). Taken together, our results suggest that the increased effectiveness in the use of WM may be related to the way training shifts attention to more phonetically diagnostic cues in synthetic speech. Before training, when listeners have additional capacity (under low WM load), they may distribute attention to irrelevant acoustic features. Prior to training, listeners may attend to synthetic speech as if it were natural speech, distributing attention across the signal in a manner that is not effective for the classification of the synthetic speech. After training, listeners can use WM more effectively because they direct attention to the acoustic properties that are most useful in recognizing speech or because they more effectively exclude the distracting or irrelevant properties (Conway et al., 2001; Lavie, 2005; Melara, Rao, & Tong, 2002). This provides an important distinction in the cognitive effects of providing training in Experiment 1, as compared with merely providing a more intelligible speech signal, as was done in Experiments 2 and 3.

Without specific training on a particular synthesizer, it seems reasonable to assume that listeners attend to the synthetic speech signal using the same distribution of attention they would use in listening to natural speech. Although the acoustic properties of higher intelligibility speech is more likely to map onto the intended phonetic categories, for both low- and high-intelligibility synthetic speech, listeners distribute attention across the speech signal as if it were natural speech. Even though high-intelligibility synthetic speech has more diagnostic cues, it still lacks the degree of cue covariation found in natural speech, still has cues that are mistaken or misleading, and still has cues that are not very diagnostic of the intended categories. Without training, listeners attend to all of these cues as if they were part of natural speech.

Lavie and colleagues (Lavie, 1995, 2001; Lavie & Cox, 1997; Lavie & Tsal, 1994) have argued that attention must always be fully allocated. According to Lavie and colleagues, the distribution of attention is obligatory; observers cannot simply choose to attend to nothing. When a primary task incurs little processing demand, observers are forced to distribute surplus attention to irrelevant (or even misleading) features or stimuli. The only way to eliminate or reduce the effect of such distracting information is then by means of higher level, WM-demanding, inhibitory processes (Lavie, 2005).

In Experiment 3, listeners in both the Votrax and CallText groups were presumably attending to those acoustic cues that they had become accustomed to through a lifetime of experience with American English. In the case of Votrax-synthesized speech (or that of any other low-intelligibility synthesizer), many of these cues were absent, misleading, or poorly identifiable. Thus, devoting a particular quantity of available resources to processing them, although unavoidable, may not have been particularly effective. In contrast, the CallText synthesizer likely provided a better assortment of such cues through patterns and contexts more like those of natural speech. In this case, devoting the same quantity of resources to processing the same set of cues resulted in better performance for the CallText speech, in which those cues were, overall, more informative. Devoting more resources to either voice should improve performance, as attention to diagnostic cues is increased. But the proportion of nondiagnostic cues that are attended to also increases, meaning that *relative* performance on the two voices should remain similar across load conditions. At low load, without training, listeners distributed attention across low-quality and high-quality synthetic speech in the same way—as if it were natural speech. However, with training, available WM was used more effectively to aid in the recognition process because training changed the distribution of attention to the speech signal, probably including some increased inhibition of irrelevant or misleading cues (cf. Melara et al., 2002). Indeed, Conway et al. (2001) showed that, when more WM capacity is available, listeners are better at filtering out distracting information (see also Lavie, 2005).

The ability to shift attention among acoustic cues may be of fundamental importance to the normal process of speech perception with natural speech. In speech perception, the basic problem of lack of invariance is that some acoustic cues in an utterance may map onto several phonological categories. This is most likely when there is a change in talker, speaking rate, or context (cf. Ladefoged & Broadbent, 1957; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Miller, 1987). In the case of context changes, there may be a momentary increase in load on WM as listeners shift attention between cues that are useful for testing between alternative phonetic hypotheses (Nusbaum & Morin, 1992). Nusbaum and Schwab (1986) suggested that learning consists of shifting attention to the cues in a signal that are more narrowly diagnostic, filtering out cues that are irrelevant or misleading, thereby decreasing the load on WM.

This kind of explanation is consistent with the way other models of perceptual attention characterize WM (e.g., Bundesen, 1990; Cowan, 1988, 1997). In these models, recognition is viewed as a process that maps stimulus features onto perceptual categories. When attention is distributed over many features, and when features activate multiple categories, the total WM demand of the system is relatively high, as is the case prior to learning synthetic speech. If learning shifts attention to a smaller set of features, or if the newly attended features are more phonetically diagnostic—or both—demands on WM should be reduced. An interesting corollary to this hypothesis is that individuals should vary in their ability to learn how to better understand unfamiliar talkers and that this variability should be related to some general measure of their WM capacity, although the direction of the relationship is open to debate. All else being equal, listeners with greater WM capacity should be able to process a greater variety of cues than can those with less capacity; but whether this should lead to better learning (because of an ability to more quickly identify diagnostic cues) or to worse learning (because of a lack of need to become more efficient; see Kane & Engle, 2000) remains to be investigated.

Training has two effects on WM use. First, learning selectively focuses attention on acoustic cues that are more diagnostic of specific phonetic categories and away from less diagnostic ones, potentially reducing the total number of attended cues. Second, because these new cues are more diagnostic, learning reduces the number of phonetic categories activated by any particular configuration of cues. As suggested by Bundesen (1990), both restricting attention to diagnostic cues and reducing the number of categories active in WM should decrease WM load.

Relationship to Other Theories

This account is compatible with studies showing other kinds of category learning (e.g., Goldstone, 1994; Livingston et al., 1998) and with studies examining changes in the use of acoustic cues in the development of speech perception (Iverson et al., 2003; Nittrouer, 1992; Nittrouer & Crowther, 1998; Nittrouer & Miller, 1997a, 1997b). It would be parsimonious to account for the development of speech perception and perceptual learning of speech and phonetic structure in adults with a single mechanism (Nusbaum & Goodman, 1994). Phonological development could be viewed as a consequence of learning from perceptuo-linguistic experience rather than from changes in the nature of the underlying perceptual mechanisms themselves.

This view of speech perception as being dynamic and dependent on cognitive mechanisms, such as selective attention and WM, is quite different from the perspective of most theories of speech perception. Motor theory (Liberman, Cooper, Harris, & MacNeilage, 1963; Liberman & Mattingly, 1985), TRACE (McClelland & Elman, 1986), LAFS (Klatt, 1979), ecological theories (Fowler & Galantucci, 2005), cohort theory (e.g., Marslen-Wilson & Welsh, 1978), feature-detector theories (Abbs & Sussman, 1971; Cooper, 1979; Stevens & Blumstein, 1981), and others (e.g., Chomsky & Miller, 1963; Pisoni & Sawusch, 1975) all employ static processing mechanisms and fixed mental representations as the starting point to establish phonetic constancy in perception of speech (Appelbaum, 1999). For example, in a TRACE-type interactive activation model of word recognition (e.g., McClelland, Mirman, & Holt, 2006), the cues available before training are relatively ambiguous, meaning they give rise to greater activation of a larger number of possible phonemes. This in turn leads to greater inhibitory competition at the phonological level and, therefore, increased processing time and decreased accuracy of recognition. However, although such models make predictions similar to those of a capacity-limitation model with regard to the effect of low- versus high-intelligibility speech, because TRACE-type models do not incorporate any limitations on overall capacity, they do not make the same predictions regarding the effects of either cognitive load or training. In a TRACE-type model, the addition of a secondary WM load should not affect processing because there is no mechanism to allow such a load to affect weights on connections between or within processing levels. Similarly, training, to the extent that it reduces cue ambiguity and, therefore, between-phoneme inhibition, would be predicted to result in an increase in accuracy and a decrease in RT, but these effects should not interact with secondary-task demands on cognitive processes. Such models are also not capable of accounting for the possibility of developing new dimensions of contrast (for a discussion, see Francis & Nusbaum, 2002).[6]

An attentionally mediated theory of phonetic learning is, however, quite compatible with recent ways of thinking about perceptual learning and categorization (see, e.g., Ahissar & Hochstein, 2002; Dinse & Merzenich, 2002; Goldstone, 2000; Schyns, Goldstone, & Thibaut, 1998). Furthermore, the proposal that phonetic features may be represented in long-term memory as abstract complexes of simpler sensory qualities generated through redistribution of attention between those qualities (Francis & Nusbaum, 2002) provides a mechanism that could account for the possibility that phonetic learning can involve the unitization and/or separation of features and, perhaps, the induction of novel features, such as has been described for visual perceptual learning (Livingston et al., 1998; Oliva & Schyns, 1997; Schyns et al., 1998). It seems reasonable to speculate that the human speech perception mechanism must incorporate some sort of adaptive learning mechanism to accomplish adult phonetic learning of new talkers and new languages, and the existence of such a mechanism could in turn prove to be a significant factor in achieving the perceptual constancy we experience daily in understanding spoken language.

The relationship between the present results and Lavie's load theory, in particular, is complex, and may not be resolvable in the context of the present results. Two issues are of particular importance in load theory—namely, whether training affects the distribution of perceptual attention or (instead) the application of available cognitive processing capacity. If training improves intelligibility by causing listeners to devote perceptual attention to fewer (but more informative) acoustic cues, this would be predicted to increase the availability of surplus capacity for distribution to irrelevant or misleading cues, thereby potentially increasing interference and, thus, RTs for correct responses. However, if training causes listeners to direct perceptual attention to a greater number of informative cues (or to cues that demanded a greater individual share of the available pool of perceptual attention), this would be predicted to result in less surplus capacity being available to be directed toward irrelevant or misleading cues, resulting in decreased interference and lower RTs. Alternatively, training may serve to alter the operation of an active filtering mechanism, in which case any training-related decrease in demands on WM capacity should be expected to result in improved efficiency of filtering, leading to a decrease in interference and an improvement in accuracy and RT in a manner that is, in principle, comparable to the theory elaborated here.

Finally, as was pointed out by an anonymous reviewer, one interesting possible consequence of the broader distribution of attention pretraining, as compared with that of attention posttraining, is that listeners might be better at identifying or remembering nonphonetic properties of the talker (e.g., indexical properties related to talker identity) before training, because their attention is initially distributed more broadly across the signal, potentially encompassing cues that are more informative of indexical rather than linguistic properties of the signal. No attempt was made to assess this possibility in the present study.

Conclusions

Results of the experiments presented here provide support for a model of speech perception in which the availability of WM capacity serves as a crucial limiting factor in performance. According to this model, speech perception proceeds via an active, hypothesis-generating mechanism, in which acoustic cues identified in the signal are used to generate hypotheses about what is being said (Nusbaum & Schwab, 1986). These hypotheses are maintained in WM and serve in part to influence top-down interrogation of the signal memory trace in order to derive more effective acoustic-cue properties, which in turn are used to further refine the phonological hypotheses in a feedback cycle limited mainly by the availability of WM (Nusbaum & Schwab, 1986; Stevens & Halle, 1967). We propose that training or experience with unfamiliar speech serves to improve the manner in which available acoustic cues are used, so that demands on WM decrease. As listeners become more familiar with the mapping between acoustic properties present in the signal and with the phonological categories those properties are intended to cue in the voice of a particular talker, they are able to direct selective attention to these acoustic cues more effectively, reducing the amount of WM capacity required to actively maintain alternative hypotheses regarding the potential interpretation of the linguistic signal. Thus, learning improves intelligibility by improving the efficiency of a resource-limited mechanism, not merely by reducing data limitations resulting from the quality of the signal.

AUTHOR NOTE

The authors thank Peter Hu for help in collecting data on intelligibility. This research was partially funded by NIH (NIDCD) Grant R03DC006811 to the first author. Correspondence concerning this article should be addressed to A. L. Francis, Purdue University, Department of Speech, Language, and Hearing Sciences, 500 Oval Dr., West Lafayette, IN 47907-2038 (e-mail: francisa@purdue.edu).

REFERENCES

Abbs, J. H., & Sussman, H. M. (1971). Neurophysiological feature detectors and speech perception: A discussion of theoretical implications. *Journal of Speech & Hearing Research*, **14**, 23-36.

Ahissar, M., & Hochstein, S. (2002). The role of attention in learning simple visual tasks. In M. Fahle & T. Poggio (Eds.), *Perceptual learning* (pp. 253-272). Cambridge, MA: MIT Press.

Appelbaum, I. (1999). The dogma of isomorphism: A case study from speech perception. *Philosophy of Science*, **66**(Suppl. 3), S250-S259.

Baddeley, A. D. (2002). The psychology of memory. In A. D. Baddeley, M. D. Kopelman, & B. A. Wilson (Eds.), *The handbook of memory disorders* (2nd ed., pp. 3-15). New York: Wiley.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47-90). New York: Academic Press.

Boothroyd, A. (1985). Evaluation of speech production of the hearing impaired: Some benefits of forced-choice testing. *Journal of Speech & Hearing Research*, **28**, 185-196.

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, **97**, 523-547.

Cepstral, Inc. (2006). David for i386-Linux (Version 4.1.2) [Software package]. Retrieved December 10, 2007. Available from www.cepstral.com/downloads/.

Chaiklin, J. B. (1955). Native American listeners' adaptation in understanding speakers with foreign dialect. *Journal of Speech & Hearing Disorders*, **20**, 165-170.

Chomsky, N., & Miller, G. A. (1963). Introduction to the formal analysis of natural languages. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 2, pp. 269-321). New York: Wiley.

Cleary, M., & Pisoni, D. B. (2001). Speech perception and spoken word recognition: Research and theory. In E. B. Goldstein (Ed.), *Blackwell handbook of perception* (pp. 499-534). Oxford: Blackwell.

Conway, A. R. A., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review*, **8**, 331-335.

Cooper, W. E. (1979). *Speech perception and production: Studies in selective adaptation*. Norwood, NJ: Ablex.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, **104**, 163-191.

Cowan, N. (1997). *Attention and memory: An integrated framework*. New York: Oxford University Press.

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

Dinse, H. R., & Merzenich, M. M. (2002). Adaptation of inputs to the somatosensory system. In M. Fahle & T. Poggio (Eds.), *Perceptual learning* (pp. 19-42). Cambridge, MA: MIT Press.

Egan, J. P. (1948). Articulation testing methods. *Laryngoscope*, **58**, 955-991.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, **11**, 19-23.

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

Fodor, J. A. (1983). *Modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

Fowler, C. A., & Galantucci, B. (2005). The relation of speech perception and speech production. In D. B. Pisoni & R. E. Remez (Eds.), *Handbook of speech perception* (pp. 633-652). Malden, MA: Blackwell.

Francis, A. L., Baldwin, K., & Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. *Perception & Psychophysics*, **62**, 1668-1680.

Francis, A. L., & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception & Performance*, **28**, 349-366.

Francis, A. L., Nusbaum, H. C., & Fenn, K. [M.] (2007). Effects of training on the acoustic–phonetic representation of synthetic speech. *Journal of Speech, Language, & Hearing Research*, **50**, 1445-1465.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, **34**, 65-89.

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

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, **123**, 178-200.

Goldstone, R. L. (2000). Unitization during category learning. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 86-112.

Goldstone, R. L., & Schyns, P. (1994). Learning new features of representation. *Proceedings of the 16th Annual Conference of the Cognitive Science Society* (pp. 974-978). Hillsdale, NJ: Erlbaum.

Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 421-433.

Hustad, K. C., & Cahill, M. A. (2003). Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology*, **12**, 198-208.

Hustad, K. C., Kent, R. D., & Beukelman, D. R. (1998). DECtalk and MacinTalk speech synthesizers: Intelligibility differences for three listener groups. *Journal of Speech, Language, & Hearing Research*, **41**, 744-752.

Ing-Simmons, N. (1994). RSYNTH: Complete speech synthesis system for UNIX [Computer software]. www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/speech/systems/rsynth/0.html

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/–/l/ to Japanese adults. *Journal of the Acoustical Society of America*, **118**, 3267-3278.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, **87**, B47-B57.

Kane, M. J., & Engle, R. W. (2000). Working memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 336-358.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Klatt, D. H. (1979). Speech perception: A model of acoustic–phonetic analysis and lexical access. *Journal of Phonetics*, **7**, 279-312.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, **87**, 820-857.

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

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception & Performance*, **21**, 451-468.

Lavie, N. (2000). Selective attention and cognitive control: Dissociating attentional functions through different types of load. In S. Monsell & J. Driver (Eds.), *Attention and performance XVIII: Control of cognitive processes* (pp. 175-194). Cambridge, MA: MIT Press.

Lavie, N. (2001). Capacity limits in selective attention: Behavioral evidence and implications for neural activity. In J. Braun, C. Koch, & J. L. Davis (Eds.), *Visual attention and cortical circuits* (pp. 49-68). Cambridge, MA: MIT Press.

Lavie, N. (2005). Distracted and confused? Selective attention under load. *Trends in Cognitive Sciences*, **9**, 75-82.

Lavie, N., & Cox, S. (1997). On the efficiency of visual selective attention: Efficient visual search leads to inefficient distractor rejection. *Psychological Science*, **8**, 395-398.

Lavie, N., Hirst, A., de Fockert, J. W., & Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General*, **133**, 339-354.

Lavie, N., & Tsal, Y. (1994). Perceptual load as a major determinant of the locus of selection in visual attention. *Perception & Psychophysics*, **56**, 183-197.

Liberman, A. M., Cooper, F. S., Harris, K. S., & MacNeilage, P. F. (1963). A motor theory of speech perception. *Proceedings of the speech communication seminar* (Vol. 2). Stockholm: Royal Institute of Technology.

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

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

Liss, J. M., Spitzer, S. M., Caviness, J. N., & Adler, C. (2002). The effects of familiarization on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria. *Journal of the Acoustical Society of America*, **112**, 3022-3030.

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 732-753.

Logan, J. S., Greene, B. G., & Pisoni, D. B. (1989). Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America*, **86**, 566-581.

Luce, P. A., Feustel, T. C., & Pisoni, D. B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*, **25**, 17-32.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, **10**, 29-63.

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

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

McCoy, S. L., Tun, P. A., Cox, L. C., Colangelo, M., Stewart, R. A., & Wingfield, A. (2005). Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech. *Quarterly Journal of Experimental Psychology*, **58A**, 22-33.

McGarr, N. S. (1983). The intelligibility of deaf speech to experienced and inexperienced listeners. *Journal of Speech & Hearing Research*, **26**, 451-458.

Melara, R. D., Rao, A., & Tong, Y. (2002). The duality of selection: Excitatory and inhibitory processes in auditory selective attention. *Journal of Experimental Psychology: Human Perception & Performance*, **28**, 279-306.

Miller, J. L. (1987). Rate-dependent processing in speech perception. In A. W. Ellis (Ed.), *Progress in the psychology of language* (Vol. 3, pp. 119-157). Hillsdale, NJ: Erlbaum.

Mirenda, P., & Beukelman, D. R. (1990). A comparison of intelligibility among natural speech and seven speech synthesizers with listeners from three age groups. *Augmentative & Alternative Communication*, **6**, 61-68.

Nittrouer, S. (1992). Age-related differences in perceptual effects of formant transitions within syllables and across syllable boundaries. *Journal of Phonetics*, **20**, 351-382.

Nittrouer, S., & Crowther, C. S. (1998). Examining the role of auditory sensitivity in the developmental weighting shift. *Journal of Speech, Language, & Hearing Research*, **41**, 809-818.

Nittrouer, S., & Miller, M. E. (1997a). Developmental weighting shifts for noise components of fricative-vowel syllables. *Journal of the Acoustical Society of America*, **102**, 572-580.

Nittrouer, S., & Miller, M. E. (1997b). Predicting developmental shifts in perceptual weighting schemes. *Journal of the Acoustical Society of America*, **101**, 2253-2266.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, **7**, 44-64.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.

Nusbaum, H. C., & Goodman, J. C. (1994). Learning to hear speech as spoken language. In J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 299-338). Cambridge, MA: MIT Press.

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

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

Nusbaum, H. C., & Pisoni, D. B. (1985). Constraints on the perception of synthetic speech generated by rule. *Behavior Research Methods, Instruments, & Computers*, **17**, 235-242.

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

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, **34**, 72-107.

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *Journal of the Acoustical Society of America*, **97**, 593-608.

Pisoni, D. B., & Sawusch, J. R. (1975). Some stages of processing in speech perception. In A. Cohen & S. G. Nooteboom (Eds.), *Structure and process in speech perception* (pp. 16-34). Berlin: Springer.

Rabbitt, P. (1991). Mild hearing loss can cause apparent memory failures which increase with age and reduce with IQ. *Acta Oto-Laryngologica*, **111**(Suppl. 476), 167-176.

Reynolds, M. E., Isaacs-Duvall, C., & Haddox, M. L. (2002). A comparison of learning curves in natural and synthesized speech comprehension. *Journal of Speech, Language, & Hearing Research*, **45**, 802-810.

Reynolds, M. E., Isaacs-Duvall, C., Sheward, B., & Rotter, M. (2000). Examination of the effects of listening practice on synthesized speech comprehension. *Augmentative & Alternative Communication*, **16**, 250-259.

Rochon, E., Caplan, D., & Waters, G. S. (1990). Short-term memory processes in patients with apraxia of speech: Implications for the nature and structure of the auditory verbal short-term memory system. *Journal of Neurolinguistics*, **5**, 237-264.

Rounsefell, S., Zucker, S. H., & Roberts, T. G. (1993). Effects of listener training on intelligibility of augmentative and alternative speech in the secondary classroom. *Education & Training in Mental Retardation*, **28**, 296-308.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime reference guide*. Pittsburgh: Psychology Software Tools.

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

Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioral & Brain Sciences*, **21**, 1-54.

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2006). Lending a helping hand to hearing: Another motor theory of speech perception. In M. A. Arbib (Ed.), *Action to language via the mirror neuron system* (pp. 250-285). Cambridge: Cambridge University Press.

Spitzer, S. M., Liss, J. M., Caviness, J. N., & Adler, C. (2000). An exploration of familiarization effects in the perception of hypokinetic and ataxic dysarthric speech. *Journal of Medical Speech-Language Pathology*, **8**, 285-293.

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

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

Tjaden, K. K., & Liss, J. M. (1995). The role of listener familiarity in the perception of dysarthric speech. *Clinical Linguistics & Phonetics*, **9**, 139-154.

NOTES

[1] See the discussion of other theories in the General Discussion of the present article.

[2] Note that Lavie and colleagues discuss the role of two kinds of capacity: perceptual and cognitive. According to their theory, greater availability of *perceptual* capacity can result in increased retention of distractors, but increased availability of *cognitive* (e.g., WM) capacity results in an improved filtering out (active inhibition) of distractors.

[3] In all cases, training was conducted over multiple days, and the posttest session and the final training session were always conducted on different days. In light of the importance of sleep and/or consolidation for perceptual learning of speech shown by previous research (e.g., Fenn, Nusbaum, & Margoliash, 2003), it is possible that the results of the present study might have been different if all training and testing has been conducted on the same day.

[4] For the purpose of comparison with older TTS systems, 20 listeners were tested using the modified rhyme test. Results showed a significant difference in percent correct responses [$t(19) = 11.13$, $p < .001$], with 93.9% correct identification for the Cepstral David voice and 70.3% for rsynth, comparable to those shown for DECtalk (96.75%) and Votrax (72.56%), respectively, as reported by Logan, Greene, and Pisoni (1989).

[5] Note that, throughout this article, all analyses with proportional data were carried out using both raw proportions and arcsine-transformed proportions, which conform better to the assumptions upon which analysis of variance is based, as suggested by Kirk (1995). In cases in which the significance of a factor differed between the transformed and untransformed tests, the transformed results are also reported. Otherwise, only the results of tests on untransformed data are reported.

[6] A notable exception to the static-mechanism assumption is the theory of "analysis by synthesis" (e.g., Stevens & Halle, 1967). However, even this model does not make any predictions regarding the role of cognitive resources, such as WM or selective attention (but for a similar proposal that does, cf. Skipper, Nusbaum, & Small, 2006).

(Manuscript received January 16, 2008;
revision accepted for publication March 15, 2009.)