3-1-2012

Rapid perceptual learning of noise-vocoded speech requires attention

Julia Jones Huyck

Ingrid Johnsrude
Western University, ijohnsru@uwo.ca

Rapid perceptual learning of noise-vocoded speech requires attention

Julia Jones Huyck[a] **and Ingrid S. Johnsrude**
Department of Psychology and Centre for Neuroscience Studies, Queen's University, 62 Arch Street, Kingston, Ontario K7L 3N6, Canada
julia.huyck@queensu.ca, ingrid.johnsrude@queensu.ca

Abstract: Humans are able to adapt to unfamiliar forms of speech (such as accented, time-compressed, or noise-vocoded speech) quite rapidly. Can such perceptual learning occur when attention is directed away from the speech signal? Here, participants were simultaneously exposed to noise-vocoded sentences, auditory distractors, and visual distractors. One group attended to the speech, listening to each sentence and reporting what they heard. Two other groups attended to either the auditory or visual distractors, performing a target-detection task. Only the attend-speech group benefited from the exposure when subsequently reporting noise-vocoded sentences. Thus, attention to noise-vocoded speech appears necessary for learning.

PACS numbers: 43.71Es, 43.71.An [SGS]
Date Received: December 7, 2011 **Date Accepted:** January 26, 2012

1. Introduction

People frequently listen to speech that is heavily accented, partially masked by background noise, or distorted due to room acoustics or electronic processing constraints (e.g., reduced bandwidth over the telephone). Fortunately, comprehension of unusual-sounding, noisy, or degraded speech improves quite rapidly, within the first few minutes of experience (see Samuel and Kraljic[1] for a review). This rapid learning has been investigated in controlled experiments in which listeners attend to speech with no distraction. In the real world, however, people often hear speech in the background while attending to another task. Does perceptual learning occur under such conditions? Are we able to improve our understanding of novel types of speech while attending elsewhere?

Perceptual learning has been demonstrated for many different types of degraded or novel-sounding speech. Naïve comprehension and perceptual learning of speech are usually measured using word report—the percentage of words that a listener is able to report correctly. In a typical perceptual-learning paradigm, word report is used to assess the intelligibility of degraded, distorted, or foreign-accented speech in listeners both before and after (and sometimes during) a training procedure that involves having them focus on understanding the degraded speech.[1–4] Learning is measured as improvement in word report between the pre- and post-training tests.

Whereas most studies of perceptual learning of speech have involved training in which attention is directed toward sentence or word comprehension, recent evidence suggests that unattended stimuli can also contribute to perceptual learning on some auditory tasks. Wright *et al.*[5] demonstrated that unattended auditory stimuli contributed to perceptual learning on a frequency-discrimination task when the unattended trials were preceded or followed by attended trials. Similarly, Seitz *et al.*[6] reported that unattended stimulus exposures led to improved perception of brief (70 ms) simulated formant transitions when the unattended exposures were temporally paired with target

[a] Author to whom correspondence should be addressed.

stimuli for an attended but unrelated task. These studies together indicate that attention may not always be necessary for auditory perceptual learning.

Apart from the perceptual learning literature, two recent neuroimaging studies suggest that people are able to process some linguistic information when they are attending to entirely different stimuli. In both investigations, speech was presented simultaneously with two non-linguistic distractor streams: one auditory and one visual. Participants were asked to attend to the speech or to one of the distractor streams during functional neuroimaging. Heinrich *et al.*[7] used this paradigm to examine whether physically interrupted stimuli that evoke the illusion of a continuous vowel still do so when attention is directed to distractor stimuli. Indeed, continuous vowels and continuity-illusion vowels showed a similar pattern of fMRI activation even when attention was directed to a distractor stream, suggesting that the continuity illusion led to perceptually complete vowels in the absence of attention. Using a similar paradigm, Wild *et al.*[8] examined how attention influences comprehension of degraded speech. Their results indicate that unattended degraded speech can be processed to some extent outside the focus of attention.

Here we use a behavioral training paradigm to examine whether attention to degraded speech is necessary for perceptual learning. In the present study, inexperienced (naïve) participants are presented with degraded sentences, simultaneously with the auditory and visual distractors used by Heinrich *et al.*[7] and Wild *et al.*[8] To assess whether unattended degraded speech contributes to learning, we compare performance among participants who attend to the degraded speech, those who attend to the visual or auditory distractors (with degraded speech in the background), and controls who are not exposed to the degraded speech or to distractor tasks.

2. Methods

We tested 72 Queen's University students (57 females) between 18 and 25 years of age [mean = 19 years old, standard deviation (SD) = 1.4 years]. All participants were recruited through email advertisement and the Queen's Psychology 100 Subject Pool. They all reported that English was their native language (12 participants indicated that they had two native languages and 25 additional participants were fluent in at least one language other than English; however, these linguistic differences did not appear to influence the results). The subjects had normal self-reported hearing, normal or corrected-to-normal vision, and no known attentional or language processing impairments. This study was cleared by the Queen's University General Research Ethics Board, and written informed consent was received from all subjects.

The experiment was organized into three phases [Fig. 1(a)]. Four groups of participants ($n = 18$ per group) completed naïve testing (5 trials) and post-training testing (20 trials). During these tests, spectrally degraded (noise-vocoded; NV) sentences were presented without any distractors and participants verbally reported as much of each sentence as they could understand. Between the 2 tests, 3 of the 4 groups were trained using 15 trials, each consisting of simultaneous exposure to NV sentences, auditory distractors, and visual distractors [see Fig. 1(b)]. One group attended to the speech, performing the same word-report task as in the naïve and post-training testing. The other two groups attended to either the auditory or visual distractors, performing a target detection task. The fourth (control) group played an unrelated silent video game for 3 min (the approximate duration of the training). No feedback was provided for any task during training or testing.

Before the naïve-testing phase, participants practiced each of the tasks in isolation (i.e., with only one type of stimulus presented at a time). All participants completed 10 trials of the visual distractor task, followed by 10 trials of the auditory distractor task and 2 trials of the speech comprehension task. These trials served to familiarize the participants with the tasks and stimuli and to ensure that they understood the task requirements. To minimize the potential for perceptual learning of speech

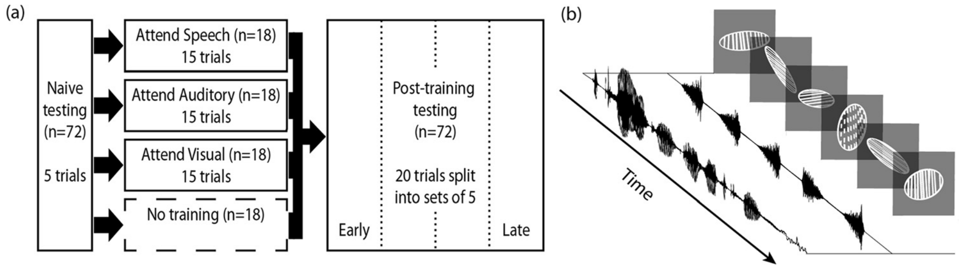

Fig. 1. (a) Schematic diagram of the experimental design. During each trial of the naïve and post-training testing, participants heard a degraded sentence and then repeated it back. During the training phase, the three trained groups were presented with degraded sentences, auditory distractors, and visual distractors simultaneously, but each group attended to a different stimulus set (performing sentence repetition or target detection). (b) Example of a single training trial. A noise-vocoded sentence is represented on the left, followed by the auditory (middle) and visual (right) distractor stimuli. All three stimuli were presented at the same time.

during this familiarization phase, we used noise-vocoded speech with 15 channels, which is quite easy to understand.

The speech stimuli consisted of 42 meaningful English sentences (e.g., "There were mice in the cave") recorded by a female native speaker of North American English in a single-walled sound booth using an AKG C1000S microphone and an RME Fireface 400 audio interface (sampling @ 16-bits, 44.1 kHz). All sentences were normalized with respect to root-mean-square (RMS) power in dB (as measured in Adobe Audition) prior to noise-vocoding. Forty sentences were split into eight sets matched for sentence duration (mean = 2116 ms, SD across sets = 189.1 ms), number of words per sentence (mean = 9, no variability across sets), number of syllables per sentence (mean = 11.4, SD = 0.7), and the logarithm of the sum word frequency (Thorndike and Lorge written frequency, mean = 5.5, SD = 0.1). Each sentence was noise vocoded by dividing it into six frequency bands selected to be approximately equally spaced along the basilar membrane[9] (cut offs: 50, 229, 558, 1161, 2265, 4290, and 8000 Hz). The amplitude envelope within each band was applied to band-limited noises with the same cut-off frequencies,[10] and the vocoded channels were then recombined. The remaining two sentences were noise vocoded in the same manner but with 15 frequency bands; these sentences were used only during the familiarization described previously.

The eight sentence sets were counter-balanced across participants and test times to minimize item effects. Importantly, three sentence sets were counter-balanced by participants across three time-points: naïve testing, early post-test (sentences 1–5; at which point the trained groups had 20 sentences of prior exposure to noise-vocoded speech), and late post-test (sentences 16–20; at which point the untrained controls had 20 sentences of prior exposure). In addition, two sentence sets were counter-balanced between post-test sentences 6–10 and 11–15. Three sentence sets were used (without counter-balancing) during training, each combined with an auditory and a visual distractor sequence.

The word-report task performed during the naïve test and post-test stages, and by participants in the "attend-speech" group during training, required participants to repeat aloud as much of each sentence as they could understand, immediately after they heard it. Participants' responses were scored for the percentage of words in each sentence that were reported correctly. Words were considered correct if they (1) matched the word in the sentence exactly (with no morphological variation) and (2) were reported in the correct order, even if intervening words were incorrect.[2,3]

The auditory distractors were amplitude modulated noise bursts (bandwidth: 1 kHz, center frequency: 4.5–5.5 kHz, duration: 400 ms). The auditory target sounded like it was "departing," having a relatively short onset ramp (50 ms) and a long offset ramp (350 ms). In contrast, the non-targets had long onset ramps (350 ms) and short offset ramps (50 ms). Each trial included between 3 and 6 stimuli, separated by a variable inter-stimulus interval (220–380 ms of silence). The number of stimuli in a sequence was chosen so that the total series duration matched the duration of the sentence that was presented on that trial [e.g., Fig. 1(b)]. Participants in the "attend-auditory" group indicated at the end of each trial whether the stimulus sequence contained a target. The "departing" target was present in 8 of the 15 trials and was never the first stimulus of the trial. The auditory distractors and noise-vocoded sentences were equated for loudness based on Glasberg and Moore's time-varying loudness model[11] and the composite stimulus was presented at an average level of ~50 dB sound pressure level (SPL) diotically over headphones.

The visual distractors were a series of rotating and stretching white cross-hatched ellipses presented on a black background. The visual target was an ellipse with broken, rather than solid, lines. The stimulus changed once every 200 ms, and the number of stimuli in a sequence was once again matched to the sentence length [e.g., Fig. 1(b)]. A "broken" target was present in 8 out of 15 trials, always within ±1000 ms of the midpoint of the accompanying sentence. Participants in the "attend-visual" group indicated at the end of each trial whether the stimulus sequence contained a target. Performance on each target detection task was assessed using a signal-detection theory measure of sensitivity (d').

Word report data were analyzed using mixed-design analysis of variance (ANOVA) with time (3 levels: naïve; early post, and late post) as a within-subjects factor and training condition (4 levels: attend speech, attend auditory distractor, attend visual distractor; control) as a between-subjects factor. Analyses were performed by subjects and items. For the analysis by subjects, scores were averaged across sentences, within participants. This type of analysis is not very sensitive if variability among items (sentences) is large. For the analysis by items, scores were averaged across subjects within the same experimental group, within items. This type of analysis is not very sensitive if variability among subjects is large. Results were deemed significant only when there was agreement between the two analyses ($\alpha = 0.05$). The Sidak adjustment was used for *post hoc* comparisons. The word-report scores were not arcsine transformed prior to conducting the statistical analyses; however, the statistical conclusions did not change when this transformation was used.

3. Results

The attend-visual and attend-auditory groups successfully directed their attention to the target-detection task during training, as indicated by excellent sensitivity on both tasks (visual distractor task: $d' = 3.668$ approximately; auditory distractor task: $d' = 2.202$).

The percentage of words reported correctly (Fig. 2) increased across test times (naïve, early post, late post), as indicated by a significant main effect of time ($p_{\text{subjects}} < 0.001$, $p_{\text{items}} < 0.001$). However, the effect of training differed among the groups: there was a significant 4-group (3 trained groups and 1 control group) by 3-time-point interaction ($p_{\text{subjects}} = 0.005$, $p_{\text{items}} = 0.033$). Note that the data from post-training sentences 6–10 and 11–15 are included in Fig. 2 for completeness but that these data were not included in the statistical analyses because of the way the sentence sets were counter-balanced.

Only the group that attended to the degraded speech benefited from the training. Word-report performance did not differ among the four groups during naïve testing (*post hoc* simple effect of group and paired comparisons: all $p\text{s} \geq 0.237$), but did differ immediately after training (early post; all $p\text{s} \leq 0.006$). Specifically, at the early post-test, the attend-speech group performed better than the untrained controls (all $p\text{s} \leq 0.006$) and the attend-auditory group (all $p\text{s} \leq 0.023$). The attend-speech group also showed a trend toward better performance after training than the attend-visual group ($p_{\text{subjects}} = 0.080$, $p_{\text{items}} = 0.026$). These group differences resolved with additional

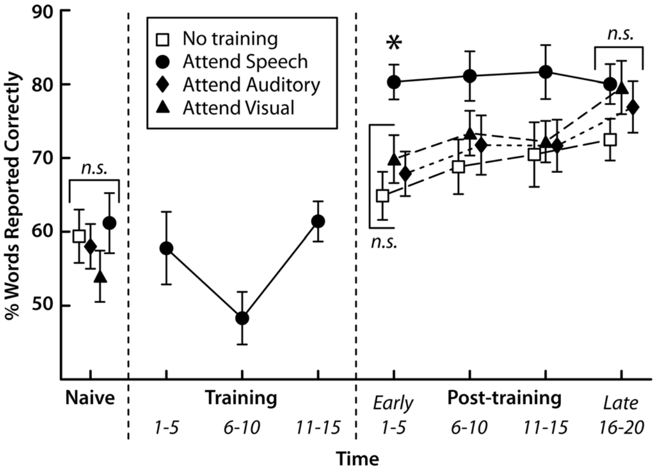

Fig. 2. Group performance on the word-report task. Mean word-report scores for each five-sentence bin. Data are shown separately for the controls (open squares); the group that attended to the speech during training (filled circles); the group that attended to the auditory distractors (filled diamonds); and the group that attended to the visual distractors (filled triangles). No statistical analyses were performed on the data from post-training sentences 6–15 because the sentence sets differed from those counter-balanced across the other test times. Error bars indicate ± one standard error.

training: performance did not differ among the four groups on the late post-test ($p_{\text{subjects}} = 0.266$, $p_{\text{items}} = 0.070$).

Unlike the group that attended to the speech, the attend-auditory and attend-visual groups performed no better than controls (and no differently from one another) immediately after training (early post; all $ps \geq 0.789$). However, because the attend-visual group *appeared* to improve slightly more than controls (see Fig. 2) and was not *significantly* poorer than the attend-speech group after training (by our criteria), we performed an additional analysis to rule out the possibility that the attend-visual group might have demonstrated some training-induced learning on the speech task. According to a 3 (groups) ×2 (time points) ANOVA, the attend-visual, attend-auditory, and control groups performed better on post-test sentences 1–5 than on the naïve test sentences (main effect of time: $p_{\text{subjects}} < 0.001$, $p_{\text{items}} < 0.001$), but they did not differ from one another overall (main effect of group: $p_{\text{subjects}} = 0.960$, $p_{\text{items}} = 0.851$) in the amount of this improvement (interaction: $p_{\text{subjects}} = 0.076$, $p_{\text{items}} = 0.138$). Thus, whatever learning was evident in these three groups at the beginning of the post-test probably resulted from the experience with the materials during naïve testing, and not from the training.

Surprisingly, even though the participants in the attend-speech group showed evidence of training-induced learning at the early post-test time point, word-report performance did not appear to improve during the training phase itself (Fig. 2, middle). Performance during the last five sentences of training was virtually identical to that during naïve testing (61.32% vs 61.21% correct) and was not significantly better than at the beginning of training (paired t-test for training sentences 1–5 vs 11–15: $p = 0.383$).

4. Discussion

The present results suggest that attention to degraded speech is necessary for learning. Only the group that attended to the speech during training demonstrated learning attributable to training. In a rehabilitative context, this means that passive listening to speech (such as half-listening to a TV program while doing a Sudoku) may not be sufficient to yield learning in individuals with poor speech comprehension due to aging, auditory processing deficits, or hearing loss. Moreover, individual differences in attentional processing may have a large impact on learning outcomes in real-world environments. The results may have particular relevance to rehabilitation of cochlear-implant patients; although it is a limited model,[10] noise vocoding bears key similarities to the way sound is transduced through a cochlear implant in that most of the spectral information is removed while the broad temporal structure is left intact.[10]

The attend-speech group reached maximal performance on the word report task during the early post-training test, with no improvement thereafter. In contrast, the other three groups improved throughout the post-training testing and demonstrated equivalent performance to the attend-speech group during the late post-training test. These results suggest that, if attention is directed toward the speech stimuli, brief training is equally effective at improving comprehension of noise-vocoded speech when there are auditory and visual distractors present (but not attended) as when the speech is presented alone.

The data also demonstrate that learning may occur without evident behavioral improvements. The attend-speech group reached asymptotic word-report performance immediately after training despite no change in performance during the training phase itself. This was unexpected, but it is possible that the competing stimuli present during the training phase may have increased processing demands (or otherwise altered behavior), perhaps masking improvement on the word-report task. Performance on the speech task in the presence of distractors may eventually have improved with additional practice. Nevertheless, from a practical standpoint, the observation that learning can be occurring even when behavior does not appear to be changing should serve as a caution to researchers and clinicians who are assessing treatment efficacy. To differentiate between ineffective and effective behavioral training, it may be necessary to assess trained performance in a controlled environment with few distractions.

It is unclear whether training benefited the attend-speech group because they performed the comprehension task with the speech stimuli or because they simply attended to the speech stimuli. It has previously been reported that perceptual learning can occur even when subjects simply listen to noise-vocoded speech without an explicit task (albeit with a slightly different trial structure than in the present study).[3] Although performing a task may help in attending to degraded speech, the task itself may not be as important as the attention.

Participants who performed the distractor tasks may have failed to benefit from their exposure to noise-vocoded speech during training for several reasons. They may simply have been attending elsewhere, may have been actively suppressing the speech signal, or the attention to and performance of the distractor task may have interfered with learning on the speech task in some other way.[12] Future behavioral and/or neuroimaging studies may help tease apart these possibilities. Alternatively, it is possible that the word report task was not sufficiently sensitive to detect small improvements in the attend-auditory and attend-visual groups relative to controls.

Finally, the noise-vocoded sentences may not have been intelligible enough to be processed as language when they were unattended. Wild *et al.*[8] demonstrated that post-test recognition memory for highly intelligible degraded speech was good regardless of the focus of attention. However, memory for low-intelligibility speech was much better when attention was focused on the speech material than when it was focused on auditory or visual distractors. Because speech intelligibility depends not only on signal quality (as described above) but also on listener experience (as in the present study), these results raise the possibility that the extent to which unattended degraded speech can be processed may increase over the course of an effective training regimen.

Acknowledgments

Conor Wild provided assistance with programming. Kirsty Chase and Kevin Shabahan helped with data collection and scoring. This work was supported by the Canadian Natural Sciences and Engineering Research Council through a Discovery Grant and EWR Steacie supplement to I.S.J.

References and links

[1] A. G. Samuel and T. Kraljic, "Perceptual learning for speech," Atten. Percept. Psychophys. **71**, 1207–1218 (2009).

[2] M. H. Davis and I. S. Johnsrude, "Hearing speech sounds: Top-down influences on the interface between audition and speech perception," Hear. Res. **229**, 132–157 (2007).

[3] M. H. Davis, I. S. Johnsrude, A. Hervais-Adelman, K. Taylor, and C. J. McGettigan, "Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences," J. Exp. Psychol. Gen. **134**, 222–241 (2005).

[4] A. G. Hervais-Adelman, M. H. Davis, I. S. Johnsrude, K. J. Taylor, and R. P. Carlyon, "Generalization of perceptual learning of vocoded speech," J. Exp. Psychol. Hum. Percept. Perform. **37**, 283–293 (2011).

[5] B. A. Wright, A. T. Sabin, Y. Zhang, N. Marrone, and M. B. Fitzgerald, "Enhancing perceptual learning by combining practice with periods of additional sensory stimulation," J. Neurosci. **30**, 12868–12877 (2010).

[6] A. Seitz, A. Protopapas, Y. Tsushima, E. L. Vlahou, S. Gori, S. Grossberg, and T. Watanabe, "Unattended exposure to components of speech sounds yields same benefits as explicit auditory training," Cognition **115**, 435–443 (2010).

[7] A. Heinrich, R. P. Carlyon, M. H. Davis, and I. S. Johnsrude, "The continuity illusion does not depend on attentional state: fMRI evidence from illusory vowels," J. Cogn Neurosci. **23**, 2675–2689 (2011).

[8] C. Wild, A. Yusuf, D. Wilson, J. Peele, M. Davis, and I. Johnsrude, "The neural system supporting the enhancement by attention of the processing of degraded speech," *Abstracts of the Proceedings of the 34th Annual Mid-Winter Meeting of the Association for Research in Otolaryngology* (2011). (A) (available online to ARO members or by e-mailing C. Wild at conorwild@gmail.com).

[9] D. D. Greenwood, "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605 (1990).

[10] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," Science **270**, 303–304 (1995).

[11] B. R. Glasberg and B. C. J. Moore, "A model of loudness applicable to time-varying sounds," J. Audio. Eng. Soc. **50**, 331–342 (2002).

[12] H. Choi, A. R. Seitz, and T. Watanabe, "When attention interrupts learning: inhibitory effects of attention on TIPL," Vision Res. **49**, 2586–2590 (2009).