Auditory attention strategy depends on target linguistic properties and spatial configuration[a)]

Daniel R. McCloy
Department of Speech and Hearing Sciences & Institute for Learning and Brain Sciences, University of Washington, 1715 NE Columbia Rd., Box 357988, Seattle, WA, 98195-7988

Adrian KC Lee[b)]
Department of Speech and Hearing Sciences & Institute for Learning and Brain Sciences, University of Washington, 1715 NE Columbia Rd., Box 357988, Seattle, WA, 98195-7988

May 16, 2015

Running title: Auditory spatial attention and linguistics

[a)]Portions of the research described here were previously presented at the 166th and 167th Meetings of the Acoustical Society of America and the 4th Gordon Research Conference on the Auditory System.

[b)]Author to whom correspondence should be addressed. Electronic mail: akclee@uw.edu

ABSTRACT

Whether crossing a busy intersection or attending a large dinner party, listeners sometimes need to attend multiple spatially distributed sound sources or streams concurrently. How they achieve this is not clear — some studies suggest that listeners cannot truly simultaneously attend to separate streams, but instead combine attention switching with short-term memory to achieve something resembling divided attention. This paper presents two oddball detection experiments designed to investigate whether directing attention to phonetic versus semantic properties of the attended speech impacts listeners' ability to divide their auditory attention across spatial locations. Each experiment uses four spatially distinct streams of monosyllabic words, variation in cue type (providing phonetic or semantic information), and requiring attention to one or two locations. A rapid button-press response paradigm is employed to minimize the role of short-term memory in performing the task. Results show that differences in the spatial configuration of attended and unattended streams interact with linguistic properties of the speech streams to impact performance. Additionally, listeners may leverage phonetic information to make oddball detection judgments even when oddballs are semantically defined. Both of these effects appear to be mediated by the overall complexity of the acoustic scene.

PACS Numbers: 43.66.Ba, 43.71.Sy, 43.66.Qp

I. INTRODUCTION

In complex auditory environments, it is sometimes desirable for listeners to divide their attention among multiple sound sources or streams, such as when attempting to follow two conversations at a dinner party. Recent psychoacoustic studies have investigated auditory divided attention, but much of the research thus far has relied on the Coordinate Response Measure (CRM) corpus (Bolia *et al.*, 2000). Because CRM sentences have uniform sentence content and a closed response set, experiments using the CRM corpus allow guessing based on incomplete phoneme perception, and do not represent the complex semantic relationships that are present in real speech (see Hafter *et al.*, 2013, for a similar critique).

In addition, recent neurolinguistic literature favors a dual-stream model of cortical speech processing in which neural signals diverge into separate dorsal and ventral streams specialized for extracting phonetic and semantic information, respectively (Hickok and Poeppel, 2004, 2007). This raises the possibility that attentional abilities and limitations may differ depending on whether a listener's task primarily relies on phonetic or semantic information in speech. In particular, the extent to which a listener has access to information from multiple speech streams may vary depending on whether their attention is directed primarily or exclusively toward speech sound or speech meaning. The experiments described here begin to address these issues by using a novel paradigm to examine how directing attention to different linguistic properties of the speech streams affects listeners' ability to direct their attention spatially.

II. BACKGROUND

The early discovery that listeners can become aware of information from apparently unattended auditory streams (Cherry, 1953) has inspired a rich literature on auditory selective attention. Most of this research has involved dichotic shadowing tasks, in which different stimuli are delivered to each ear and listeners continously repeat aloud ("shadow") the speech in the attended ear. The picture that emerges from such studies is that listeners are generally aware of

gross acoustic features of the unattended stimulus (such as speech versus tone, or male versus female voice, cf. Cherry, 1953; Lawson, 1966), but are not aware of more subtle aspects of the unattended signal (such as its semantic content, the language being spoken, or whether it has been time-reversed, cf. Cherry, 1953; Wood and Cowan, 1995b). However, certain types of information (such as the listener's own name) may draw attention to the rejected stream (Moray, 1959; Wood and Cowan, 1995a), and listeners may even be aware of irrelevant information in the rejected channel if they are queried about it soon enough after its presentation (Norman, 1969; Glucksberg and Cowen, 1970). There have also been some studies suggesting that the unattended speech is processed at least enough to cause priming effects, even when listeners are not consciously aware of the content in the rejected channel (Eich, 1984; Wood *et al.*, 1997; Dupoux *et al.*, 2003; Rivenez *et al.*, 2006).

These findings regarding auditory selective attention are relevant to questions regarding divided attention because there is some question as to whether true auditory divided attention is even possible. Some theories of perception (e.g., Broadbent, 1958) assert that perception must ultimately involve a stage of serial information processing (the so-called "single channel" hypothesis). Such theories require that, in a divided attention task, one of the two "simultaneously" attended streams must be stored in memory until after information from the first stream has been processed. While the single channel hypothesis appears to be inaccurate on a domain-general view of attention (cf. Allport *et al.*, 1972), it remains controversial whether it is possible to divide one's *auditory* attention across two or more distinct *auditory* objects or streams. Establishing the processing of unattended information (as the priming studies seem to do) may provide evidence against the single channel hypothesis within the auditory domain, instead favoring an "attenuation model" (Treisman, 1960) in which unattended stimuli are processed to varying degrees. Still, such studies do not establish the ability to *consciously attend* to multiple streams.

Current psychological models of auditory spatial attention favor a "spotlight" metaphor borrowed from the literature on visual attention (Eriksen and Hoffman, 1972; Eriksen and

St. James, 1986), but it is thus far unclear how this auditory spotlight operates. Candidate models of divided auditory attention proposed by Best and colleagues (2006) include a single, broadened spotlight similar to the visual "zoom lens" model due to Eriksen and St. James (1986); a single, narrow spotlight that rapidly switches between attended locations; and dual (or multiple) narrow spotlights deployed in parallel. Of these, the "parallel spotlights" model is the only one allowing for truly simultaneous attention to distinct auditory objects.

In their study of divided attention using the CRM corpus, Best and colleagues (2006) could not fully disambiguate which of these models best explained their findings. They reported conflicting results from experiments using different-band tone-vocoded speech (for which listener performance peaked around 90° separation) and similar experiments using natural speech (for which performance was best at smaller separation angles). Analysis of listener errors in the natural speech task was suggestive of a "single spotlight with rapid switching" model, but the existence of errors that were not consistent with this model prevented the authors from drawing any definitive conclusions. Additionally, their results were suggestive of a listener strategy that prioritized attention to the left-hand stream, and relied on introspection of something like "echoic memory" or "sensory traces" of the right-hand stream to accomplish the divided attention task (the prioritizing of the left-hand stream was prompted by the task instructions to report keywords from the left stream first).

A related study also using the CRM corpus investigated the effect of differences in stream level on the benefit of spatial separation in a divided attention task (Ihlefeld and Shinn-Cunningham, 2008). Using different-band tone-vocoded speech with one stream at a fixed intensity and the other varying across trials from −40 dB to 0 dB (relative to the fixed-intensity stream), they found a pattern of performance and errors suggesting that listeners were prioritizing attention to the quieter (varying-intensity but fixed call-sign) stream, and reporting keywords from the fixed-intensity (but varying call-sign) stream based on recall from temporary storage. In particular, they found that spatial separation of the streams improved performance on the prioritized stream only, which they interpret as evidence for its being actively attended.

However, the task design in this experiment (and in the experiments of Best *et al.*, 2006) necessitated greater reliance on memory for whichever stream was "lower priority," both because listeners had to delay any response until after the trial stimulus ended, and because they tended to report the higher priority stream first (further delaying report of keywords from the lower priority stream). For this reason, the results of these studies cannot be taken as conclusive evidence that true simultaneous attention to distinct auditory objects or streams is an impossibility.

Using similar stimuli presented dichotically, Gallun and colleagues (2007) found that dividing attention across two streams entailed a decrease in performance (compared to selective attention trials with the same stimuli) when the task in both ears involved keyword identification. Although the dichotic stimulus presentation prevents drawing any conclusions about the spatial profile of auditory attention from this study, the authors did find an interesting dual-task effect: listening for keywords in both streams (and only afterward being cued which stream to report) yielded a performance cost, compared to a task that required merely detecting the presence or absence of speech in noise in one ear while reporting keywords in speech delivered to the other ear. A follow-up experiment with keywords only (no carrier sentence) yielded the same results, which the authors interpreted as evidence against a within-trial attention-switching strategy (i.e., listeners were not first detecting the presence of speech in one ear, then switching attention to the other ear for keyword identification). This latter finding is at least compatible with simultaneous attention to both streams, though certainly not definitive proof of it, while the performance reduction in the dual-keyword-reporting task is harder to reconcile.

A more recent study by Hafter and colleagues (2013) examined divided spatial attention using read passages from short stories. The three concurrent stories were interrupted to ask participants questions that relied on either phonetic or semantic information from one of the streams, that had occurred at varying latencies before the interruption. They found that when listeners knew that only 60% of post-trial questions would relate to a talker at the midline (the remaining 40% split between two flanking talkers), the spatial release from masking seen in

selective attention trials disappeared and a trend emerged for semantic information from the flanking talkers to be more readily available when the flankers were spatially near (±7.5°) rather than spatially distant (±60°). The same trend was seen in trials where the probe question relied on phonetic information in the flanking talkers' speech. Hafter and colleagues interpret this as most consistent with a "single broadened spotlight" model of auditory spatial attention rather than a switching strategy, though the authors admit that their evidence is far from conclusive. Their results could also be interpreted as consistent with a switching strategy, with the added assumption that switching between streams with a larger separation angle is more costly than switching between more closely located streams.

Taken together, the results of these studies present a picture of divided auditory spatial attention that is far from clear. The somewhat incongruous results may reflect between-listener variation with regard to which strategies are preferred or most effective (regardless of task type). If such variation were present in the population, it could manifest as conflicting results depending on the particular listeners sampled in each study. The differing results may also mean that listeners have access to more than one strategy, and details of the task design might favor one strategy over another. In that case, differences in task design between studies would be the cause of the different results. Of course, the differences in task design among these studies are well-motivated, in that each is trying to model the classic "cocktail party" problem (Cherry, 1953) as realistically as possible while minimizing or controlling confounding factors — especially the level of energetic masking that arises in a multi-stream listening environment. In the three studies by Best and colleagues, Ihlefeld and Shinn-Cunningham, and Gallun and colleagues mentioned above, priority was given to mimization of energetic masking, which was achieved by synthesizing the speech through tone vocoding using different frequency bands for the different speech streams (at the expense of naturalness of the stimuli). In contrast, Hafter and colleagues prioritized faithfulness to the natural setting by presenting streams of unprocessed running speech from different talkers at different spatial locations, without explicitly controlling for issues related to energetic masking.

The present study takes a different approach, and attempts to strike a middle ground between the approaches described above. In these experiments, trials comprise speech streams built from sequences of monosyllabic words. This allows precise control of both the lexical properties of the words chosen, as well as their relative timing within each trial. Controlling the timing helps to minimize temporal overlap (and thus energetic masking) and maximize word onset cues, while controlling lexical properties (such as word frequency, phonotactic probability, and phonological neighborhood density) helps remove variation in the ease of word comprehension that can arise in both closed-set tasks and in running speech. In this paradigm, the strings of monosyllabic words are united by either their phonetic or their semantic properties; these within-stream linguistic relationships can be thought of as a proxy for the relationships among words in a typical sentence of natural speech. Moreover, by using the same talker for all streams and monotonizing the words to remove any extraneous pitch cues, these experiments ensure that listeners are relying only on acoustic segmental cues to identify the words, and have only spatial cues (primarily interaural time and level differences) to rely on in making their judgments of spatial location.

Importantly, the behavioral response in these experiments is a rapid button press during the trial (when an oddball item is detected), rather than a trial-end response such as keyword repetition. Consequently, both the necessity of relying on memory and its efficacy at improving performance are reduced. For this reason, we believe the results of these experiments reflect listeners' (near)-realtime access to information from different speech streams, rather than their ability to use memory to compensate for a serial processing bottleneck in tasks that require parallel processing of incoming stimuli.

III. GENERAL METHODS

Experiments 1 and 2 both involve four streams of monosyllabic words, with distinct spatial origins for each stream simulated by processing with pseudo-anechoic head-related transfer

functions (HRTFs) at ±15° and ±60° azimuth in the horizontal plane (the recording of these HRTFs is described in Shinn-Cunningham *et al.*, 2005). Each stream comprised 12 words, for a total of 48 words per trial. On each trial, 3–4 oddball words were chosen to replace existing words in the trial; oddballs were defined for the participant as any word not matching the category of the spatial stream in which they occurred (in semantic trials) or any word other than the repeated base word of the stream (in phonetic trials). Prior to and throughout each trial, one or two of the four streams was cued visually as "to-be-attended," and listeners responded by button press to oddball stimuli in the to-be-attended stream(s) while trying to ignore oddballs in other streams.

A. Participants

All participants in Experiments 1 and 2 had normal audiometric thresholds (20 dB HL or better at octave frequencies from 250 Hz to 8 kHz) and were compensated at an hourly rate. All subjects gave informed consent to participate in the study as overseen by the University of Washington Institutional Review Board.

B. Stimuli

The words used to construct the trials were recordings from a single talker, normalized to have the same root-mean-square (RMS) amplitude and monotonized to the talker's mean f_0 using the Praat implementation of the PSOLA™ algorithm (Boersma and Weenink, 2014; Moulines and Charpentier, 1990). This was done to limit stream segregation cues to spatial cues only (primarily interaural time and level differences) generated by convolution with the HRTFs. Across streams there was a consistent delay of 250 ms from the onset of one word to the onset of the following word, but streams were interdigitated such that within-stream onset-to-onset delay varied between 750 and 1750 ms. This eliminated any rhythmic regularity to the timing of words in a given stream, so that listeners could not form temporal expectations about when the next word in a to-be-attended stream might occur. Word durations ranged from 336 to 783 ms

(mean 557 ms) so there was some temporal overlap between sequential word tokens (though the constraints on interdigitation ensured there was rarely overlap between two tokens in the same spatial stream). A diagram of a typical trial is illustrated in Figure 1.

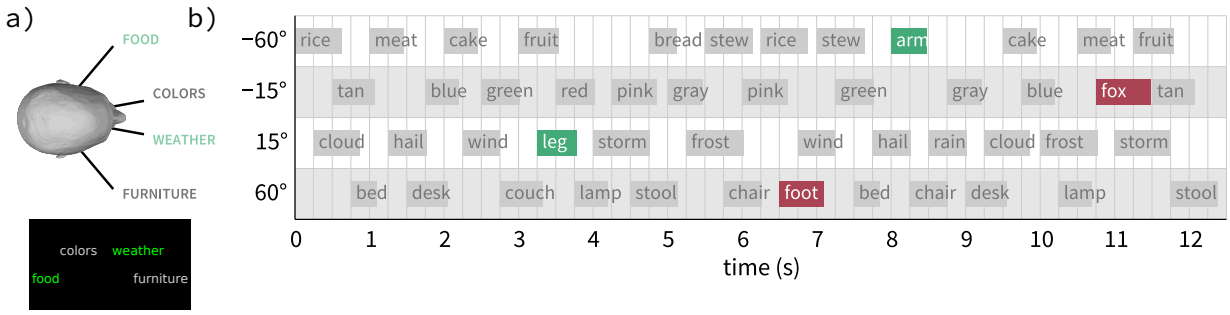

Figure 1: (Color online) Diagrams of trial structure. (a) Diagram of the spatial location of the four streams (top) and a corresponding screenshot of the visual prompt (bottom) showing how the four spatial locations were represented on screen, and the color cueing which streams are to-be-attended. (b) Schematic trial time course showing a semantic trial in the spatially non-adjacent, divided attention condition. To-be-attended streams have light backgrounds, to-be-ignored streams have darker backgrounds. The width of the small rectangles correspond to actual word durations; rectangles for oddball words have light text on darker backgrounds. (The information in this figure may not be properly conveyed in grayscale.)

In creating each trial, the intensities of the four streams were equated by finding the mean RMS amplitude across the left and right channels for each stream, and using these mean values to calculate a scaling factor to be applied to each stream. Each stream's scaling factor was applied to both left and right channels. This procedure equated the overall energy at the ears associated with each stream, while preserving the relative intensities of the channels and thus preserving the interaural level difference cues. The stereo waveforms were then summed across the four spatial locations and scaled to a presentation level of 65 dB SPL to generate the final trial waveform. For any given word in an attended stream, energetic masking due to (parts of) words in the other streams was fairly low. Mean signal-to-masker ratio was +2.97 dB with a standard deviation of 4.13 dB (calculated monaurally at the better ear for the attended word).

C. Procedure

The auditory stimuli and visual prompts were presented, and participant responses collected, using expyfun software (Larson *et al.*, 2014). A five-step training procedure ensured mastery of all aspects of the task (identify targets, attend single streams, attend multiple streams, ignore streams, etc). Each trial included a persistent visual prompt beginning 2.5 s before the audio onset, cueing participants to the correspondence between spatial location and stream content. In semantic trials, this cue was the name of a category to which all non-oddball words in that stream belonged. In phonetic trials, each stream comprised repetitions of a single "base" word with occasional oddballs, so the visual cue was the text of the non-oddball "base" word in that stream. The color of the words in the visual prompt indicated which stream(s) were designated as to-be-attended (green) and to-be-ignored (grey). Stimuli were delivered over insert earphones in a sound-attenuated booth via a Tucker Davis Technologies RP2 realtime processor at a presentation level of 65 dB SPL. Throughout each block there was also a 40 dB SPL white noise masker; this was included for consistency with future neuroimaging studies, and is due to the acoustic conditions in the neuroimaging suite. The noise masker had signal polarity flipped between left and right channels to create a diffuse noise image, so that the noise would not have a perceivably distinct spatial location (which could potentially interfere with spatial stream selection).

Stimuli were grouped into blocks of 4–8 minutes; participants were encouraged to take breaks between blocks at their discretion. Prior to each experiment, participants performed a computerized categorization task to ensure there was no uncertainty about which semantic category each word belonged to. This task (along with the five-step training) also ensured that listeners were "overtrained" on the words to be used, further minimizing any differences between the familiarity of words across categories.

D. Analysis

The response window for each word was defined as the span from 250–1250 ms after word onset. Because the across-stream presentation rate was one word every 250 ms, a given button press could conceivably be assigned to the timing slot of any of four sequentially presented words (fewer at the beginnings and ends of trials). By design, oddballs were far enough separated in time such that no two oddballs could ever both be candidates for attribution of the same button press. Assignment of a button press to a particular word was done by first checking if any of the four candidate words were oddballs; if so the response was assumed to reflect detection of the oddball, the button press was assigned to that oddball's timing slot, and response time was calculated relative to the onset of the oddball word. Otherwise, the response was deemed a "stray press" and was arbitrarily attributed to earliest-occurring non-oddball word for which the button press fell within its response window. This biased the response times for stray presses to be rather high (i.e., 1000–1250 ms), however, since stray presses were merely being counted as false alarms without further analysis of the response latency, the bias was of no consequence.

The experimental design admits of more than one possible calculation for listener sensitivity (d'). For example, since there are 12 words per stream and 4 streams, a trial with 3 targets in which the listener responded perfectly could be said to have 45 correct rejections. Including all non-target words as possible correct rejections — even those in unattended streams — has the effect of inflating d' values across the board, potentially obscuring differences between subjects or conditions. It is also unrealistic given our stimulus design, since we ensured that each 1-second stimulus frame contained 1 word onset from each spatial location, and there were never targets from both attended streams within the same 1-second frame, so there was never a situation in which two or more button presses were warranted within 1 second of each other. For these reasons, we chose to calculate d' based on the twelve 1-second frames of each trial; in other words, a trial with 3 targets would have up to 9 correct rejections instead of 45.

To calculate chance performance, we make the simplifying assumption that if a target and a button press occurred in the same 1-second frame, the button press was in response to the target

and should count as a hit regardless of which came first (of course, for the actual participant data we compared time of target onset to time of button press, as described above). This assumption has the effect of slightly raising chance performance, and as such can be seen as a conservative assumption as it sets a higher bar for assessing whether the performance of actual subjects is non-random. Given this assumption, and the further assumption that the chance listener knows the mean rate of 2.5 target frames out of 12 total frames per trial, a 0.2083 probability of button press in each 1-second frame yields a d' value of 0.01 for both experiments. If we make the stronger assumption that our chance listener presses in accordance with the target rate (2.5 per trial) and is correct 50% of the time (i.e., 50% of their button presses are hits and the other 50% are false alarms), this yields chance values of 1.13 for Experiment 1 and 1.12 for Experiment 2.

E. Statistical methods

Listener sensitivity was modeled with generalized linear mixed-effects regression using the lme4 package (Bates *et al.*, 2014) in the R statistical computing environment (R Development Core Team, 2014). The statistical model for listener sensitivity was constructed to predict the probability of listener button presses at each timing slot of the trial (one timing slot per word, cf. Section III.D.). The general form of the model equation is given in (1):

(1) $Pr(Y = 1 \mid X) = \Phi(X'\beta)$

In (1), Y is the binary response (i.e., $Y = 1$ means the listener pressed the response button), Φ is the cumulative normal distribution function, β is the vector of coefficients to be estimated, and X is a design matrix describing whether the current word is a target, foil, or neither (hereafter "word type") and additional aspects of the experimental design to be used as predictors (e.g., attention to one stream or two, whether the two attended streams were spatially adjacent, etc.). Equation (1) can be reformulated to make use of the inverse of the cumulative normal distribution function (Φ^{-1}, also known as the normal quantile function or probit function), as seen in (2):

(2) $\Phi^{-1}(Pr(Y = 1 \mid X)) = X'\beta$

When X simply indicates the presence of a genuine target, the expression $Pr(Y = 1 \mid X)$ is equivalent to hit rate. In this light, the left side of Equation (2) bears a striking relationship to the first term of Equation (3), the common equation for estimating d' (cf. Macmillan and Creelman, 2005):

(3) $\Phi^{-1}(\text{hit rate}) - \Phi^{-1}(\text{false alarm rate}) = d'$

Formulating Equation (2) as a mixed-effects model (i.e., estimating contributions of both population-level characteristics and individual-level effects) models hit rate and false alarm rate as consistent within-listener but potentially varying across listeners, and subject to population-level influences (Sheu *et al.*, 2008). This design allows estimation of population-level effects (such as the effects of task design or experimental condition) to be based on data from all participants, without assuming that all participants use the same decision criterion when performing the task. Moreover, modeling response probabilities on a probit scale allows model coefficient estimates to be interpreted as d' values. In other words, because the model equation can be expressed as a sum of terms of the form $\Phi^{-1}(k)$, and d' is likewise estimable as a sum of such terms, the magnitudes of model coefficients can be interpreted as d' differences between conditions (assuming appropriate coding of the predictor variables). This modeling strategy is based on the formalizations in DeCarlo (1998) and Sheu *et al.* (2008); cf. Lawrence and Klein (2013) for a recent study using a similar approach to modeling sensitivity in an audiovisual attention task.

One additional advantage of our model design is that responses to foils are estimated separately from other types of false alarms, so the effect of experimental manipulations on both target response rate and foil response rate can be examined. However, because the equation is modeling probability of response (rather than modeling d' directly), we must be careful when interpreting the signs of the coefficient estimates. That is, a coefficient involving word.type=target will have a direct relationship with d' (i.e., positive coefficient indicates

increased target detection and thus higher d'), whereas coefficients involving word.type=foil will have an inverse relationship with d' (i.e., positive coefficient indicates increased false alarm responses to foils, thus lower d'). This inverse relationship is also apparent if one considers the subtraction in the d' estimating equation shown in (3).

IV. EXPERIMENT 1

The first experiment investigated listeners' ability to detect oddball stimuli in cued streams and ignore oddball stimuli in uncued streams, with oddballs defined based on either phonetic or semantic features of the stimuli. 60 phonetic and 120 semantic trials were presented to each subject. Within each of these two trial types, 40% were selective attention trials, 30% were divided attention trials with adjacent streams cued as to-be-attended, and 30% were divided attention trials with non-adjacent streams cued as to-be-attended.

A. Methods

1. Participants

Fourteen participants (ten female), aged 21-32 (mean 25), were recruited for Experiment 1. One male participant did not complete the experiment for personal reasons. Six participants were presented the semantic condition first; the remaining seven were presented the phonetic condition first. All participants completed the pre-experiment category familiarization task without errors, and passed the five-step training procedure.

2. Stimuli

Stimuli for the semantic condition comprised sets of six monosyllabic words in each of seven semantic categories (see Appendix). An additional twelve words were recorded for use in the phonetic condition. The semantic categories did not statistically differ in the lexical frequency

of their words, nor in the phonological neighborhood density of their words, nor the mean uniphone or biphone frequencies of their words (measures of uniphone and biphone frequency reflect phonotactic probability, or how likely a sequence of phonemes is relative to all other words in the language). Words exhibiting polysemy that might potentially place them into more than one category were excluded, as were words that formed phonological minimal pairs with words that would have been congruent with one of the other semantic categories. Statistical summaries comparing lexical properties of the semantic categories are given in Table I.

Table I: Summary of analysis of variance results for some lexical properties of the semantic categories in Experiment 1. Phonotactic probabilities were calculated using an online tool described in Vitevitch and Luce (2004); lexical frequency and neighborhood density data were drawn from Sommers (2014).

Lexical property	**Summary statistics**
lexical frequency	$F(6,35)=0.92$, $p=0.50$
phonolgical neighborhood density	$F(6,35)=0.52$, $p=0.79$
mean uniphone frequency	$F(6,35)=0.78$, $p=0.59$
mean biphone frequency	$F(6,35)=1.53$, $p=0.20$

Semantic trials were constructed by selecting four categories and assigning each category to a spatial location; the assignment of category to location was held fixed within each experimental block (30 trials per block). Order of words within each spatial stream was random. On each trial, 3–4 words were replaced with oddball words (words not matching the semantic category of the stream in which they occurred). These oddballs were drawn from any of the three semantic categories not in use during that block. Oddballs were pseudo-randomly distributed, constrained such that (a) oddballs could not be the first or last word in their stream, and (b) no oddballs occurred in sequential 1-second frames (across all streams). On each trial either 2 or 3 of the oddballs occurred in to-be-attended streams ("targets"), and 0–2 occurred in to-be-ignored streams ("foils"). Total trial duration was just over 12 seconds, yielding a block length of about 8 minutes.

In phonetic trials, four individual words (rather than categories) were selected to correspond to

each spatial location, and each spatial stream comprised 12 repetitions of the word chosen for that location. As in the semantic trials, the assignment of word to spatial location was held fixed across each block, and word onsets were distributed and constrained in the same fashion. Phonetic trial oddballs were defined to participants as any word not identical to the base word in the spatial stream in which they occurred; number and distribution of targets and foils within each trial was the same as in the semantic trials. Phonetic trial blocks contained 15 trials instead of 30, yielding a block length of about 4 minutes.

3. Procedure

In each trial the visual prompt indicated the correspondence between spatial location and stream base word (phonetic trials) or stream category (semantic trials), and a color cue denoted which stream(s) were to-be-attended. Listeners were instructed to respond to target words as quickly as possible via button press.

4. Statistical analysis

As described in Section III.E., listener sensitivity was modeled with generalized linear mixed-effects regression. The model predicted probability of listener button press based on whether the current word was a target, foil, or neither ("word type"); whether the current trial was a phonetic or semantic trial ("trial type"); and whether the listener was cued to attend to one stream, two adjacent streams, or two non-adjacent streams ("attentional configuration"). Estimation of the main effects requires five coefficients (two for each of the ternary categorical predictors "word type" and "attentional configuration," and one for the binary predictor "trial type"). Estimation of the two-way and three-way interactions adds another 12 coefficients, plus 1 for the overall intercept for a total of 18 fixed-effect coefficients (cf. the "Indicator" column in Table III). Additionally, the model includes a random intercept for participant, which estimates a variance component around the grand mean such that the model fits a (potentially) unique

intercept for each listener. The model equation is shown in schematic form in (4):

$$(4)\quad \Phi^{-1}(y_{ij}) = \beta_0 + \beta_1 W_{ti} + \beta_2 W_{fi} + \beta_3 T_i + \beta_4 C_{ai} + \beta_5 C_{si} + \ldots + S_{0j} + \epsilon_{ij}$$

In (4), y_{ij} is the outcome (button press) for word i and subject j, β_0 is the intercept term, and the other β terms are the coefficients estimated for the various predictors. W_{ti} and W_{fi} are binary indicators for whether word i was a target, foil, or neither; in other words, the ternary "word type" predictor has treatment coding and word.type=neither as baseline. T_i indicates trial type (semantic or phonetic), and has deviation coding. The coefficient estimated for T_i reflects the difference (on a d' scale) between semantic trials and phonetic trials (semantic minus phonetic). C_{si} indicates whether listeners were cued to attend to one or two streams, and C_{ai} indicates attention to two adjacent streams versus two non-adjacent streams (i.e., the ternary "attentional configuration" predictor has reverse Helmert coding). The coefficient estimated for C_{ai} reflects the difference between divided attention trials where attended streams are spatially adjacent versus spatially non-adjacent (adjacent minus non-adjacent). The coefficient estimated for C_{si} reflects the difference between selective versus divided attention trials (selective minus divided), collapsing across the adjacent/non-adjacent distinction within the divided attention trials. A summary of factor coding of the fixed effects in this model is given in Table II. The ellipsis in Equation (4) indicate that additional coefficients were estimated for the two- and three-way interactions of the fixed effects; S_{0j} is the random effect for subject j, and ϵ_{ij} is the error term. The coding of the predictor variables described above and summarized in Table II yields the following interpretation of the estimated coefficients. The experimental manipulations ("trial type" and "attentional configuration") are coded so that their coefficient estimates reflect differences between conditions (as discussed above), and their *p*-values can be interpreted straightforwardly. The "word type" coefficients target, foil, and neither (neither is the intercept term) provide estimates of baseline response levels to target, foil, and non-target non-foil words (respectively), across all trial types and attentional configurations. The model estimates probability of response, so positive coefficients indicate high probability of response, and negative coefficients indicate low probability of response. Thus after probit transformation we

Table II: Coding of indicator variables in the statistical model for Experiment 1.

Factor	Coding	Indicator	Coef. name	Value
word.type	Treatment	W_{ti}	target	1 if word is target, 0 otherwise
		W_{fi}	foil	1 if word is foil, 0 otherwise
attn.config	Helmert	C_{si}	selective	⅔ if attend one stream ⁻⅓ if attend two streams
		C_{ai}	adjacent	0 if attend one stream 0.5 if attend two adjacent streams −0.5 if attend two non-adjacent streams
trial.type	Deviation	T_i	semantic	0.5 if semantic trial, −0.5 if phonetic trial

would expect the coefficient for target to have a large positive coefficient (assuming that responses to targets were frequent), foil to have a coefficient closer to zero (either positive or negative, depending on foil response rates), and neither to have a negative coefficient (assuming that responses to non-target non-foil items were rare). Thus the intercept and the target coefficient (and possibly the foil coefficient) are expected to be highly significantly different from zero, although this is much less interesting than the significance of the differences between conditions of the experimental manipulations.

The effects of experimental manipulations are split into three groups of coefficients. The effects of manipulations on *response bias* are assessed by examining coefficients for the experimental manipulations alone and their interactions with one another (e.g., the coefficient for the model term T_i compares responses in semantic versus phonetic trials *across all words in the trial* regardless of attentional configuration or whether the words were target, foil, or neither). The effects of experimental manipulations on *target detection sensitivity* are assessed by examining estimates for the interactions between target and the experimental manipulations (e.g., the coefficient estimate for $W_{ti} : T_i$ compares responses in semantic versus phonetic trials *for target words only*). The effects of experimental manipulations on *response to foil items* are assessed by examining estimates for the interactions between foil and the experimental manipulations (e.g., the coefficient estimate for $W_{fi} : T_i$ compares responses in semantic versus phonetic trials *for*

foil words only).

Using the approach to calculating d' described in Section III.D., perfect performance across all trials in this experiment yields a d' ceiling of 6.50; the highest performance of any subject in any condition was a d' of 5.71 in the phonetic selective attention trials. As mentioned in Section III.D., chance performance for this experiment is conservatively estimated as a d' value of 1.13. The lowest performance of any subject was a d' of 1.42, in the condition with semantic target definitions and divided attention to spatially separated streams. This suggests that none of the experimental conditions were so difficult that subjects had to resort to random response strategies.

B. Results

The model summary for listener responses is seen in Table III, and corresponding barplots of d' values are shown in Figures 2 and 3. The model coefficients are grouped into coefficients giving baseline response levels, coefficients indicating differences in bias among the experimental conditions, coefficients indicating differences in response to target items, and coefficients indicating differences in response to foil items.

The variance estimated to account for differences between subjects (S_{0j} in Equation (4)) was quite small (standard deviation of 0.098 on a d' scale), suggesting that performance across subjects was quite consistent. Baseline response levels show the expected pattern: responses to targets were generally high (coefficient of 3.31), whereas response to foil items were less common (coefficient of 0.97), and responses to non-target non-foil items were quite rare (coefficient of –2.62). As discussed above, the statistical significance of these baseline predictors is not very interesting scientifically, since they merely tell us that target hit rate was much higher than 0.5, foil response rate was also higher than 0.5, and response rate to non-target non-foil items was much lower than 0.5 (0.5 corresponding to 0 via the probit transformation). Among the coefficients for response bias, only the coefficient for semantic versus phonetic condition is noteworthy; the others are either not statistically significantly different from zero,

Table III: Model summary predicting listener button presses in Experiment 1. * indicates p<0.05, ** indicates p<0.01, *** indicates p<0.001. † indicates significant coefficients (at the p<0.001 level) that are based on treatment coding; significance for these coefficients is expected and should be interpreted differently than the other coefficients (see text for explanation). SE = standard error of the coefficient estimate. Interactions between predictor levels are indicated by colons.

Indicator	Predictor name	Coef.	SE	z	p	Signif.
Baseline response levels						
	neither (Intercept)	−2.62	0.03	−78.55	<0.001	†
W_{ti}	target	3.31	0.03	112.46	<0.001	†
W_{fi}	foil	0.97	0.06	17.54	<0.001	†
Effect of manipulations on response bias						
T_i	semantic	0.42	0.04	10.78	<0.001	***
C_{si}	selective	−0.12	0.04	−2.89	0.004	**
C_{ai}	adjacent	0.02	0.05	0.45	0.650	
$T_i : C_{si}$	semantic:selective	0.18	0.09	2.13	0.033	*
$T_i : C_{ai}$	semantic:adjacent	0.05	0.09	0.56	0.576	
Effect of manipulations on response to targets						
$W_{ti} : T_i$	target:semantic	−1.44	0.06	−24.56	<0.001	***
$W_{ti} : C_{si}$	target:selective	0.79	0.06	12.26	<0.001	***
$W_{ti} : C_{ai}$	target:adjacent	0.11	0.07	1.55	0.120	
$W_{ti} : T_i : C_{si}$	target:semantic:selective	0.18	0.13	1.36	0.173	
$W_{ti} : T_i : C_{ai}$	target:semantic:adjacent	−0.23	0.14	−1.62	0.105	
Effect of manipulations on response to foils						
$W_{fi} : T_i$	foil:semantic	−0.73	0.11	−6.58	<0.001	***
$W_{fi} : C_{si}$	foil:selective	−0.28	0.12	−2.30	0.021	*
$W_{fi} : C_{ai}$	foil:adjacent	−0.58	0.14	−4.32	<0.001	***
$W_{fi} : T_i : C_{si}$	foil:semantic:selective	0.14	0.25	0.57	0.570	
$W_{fi} : T_i : C_{ai}$	foil:semantic:adjacent	0.83	0.28	2.97	0.003	**

or have magnitudes that are too small (less than 0.2 on a d' scale) to confidently interpret. Generally speaking, significant bias coefficients in this model are likely driven by responses to non-target non-foil items ("stray presses"), which may indicate random mistakes or slow responses to targets or foils that fell outside the response window. The bias toward responding more in the semantic condition is somewhat unexpected, but likely reflects the difference in stray presses between the two trial types: there were 83 such responses total across subjects in phonetic trials, versus 577 in semantic trials. This disparity is almost certainly the cause of the difference in bias between the phonetic and semantic conditions. In contrast, false alarm responses to *foil items* were similar between the conditions (63 foil responses total across subjects in phonetic trials, versus 49 in semantic trials).

Significant main effects indicate that target detection is better in phonetic trials (the coefficient for target:semantic is negative, cf. Figure 2a), though there are also more responses to foil items in phonetic trials (foil:semantic is negative). There is also a main effect for selective versus divided attention trials: target detection is better, and response to foils less likely, when attending only one stream (target:selective is positive and foil:selective is negative; cf. upper bracket in Figure 2b). The apparent main effect of spatial adjacency (foil:adjacent is positive; cf. lower bracket in Figure 2b) appears to be entirely driven by its effect in the phonetic condition, and is discussed further below.

There is also a significant interaction between attentional configuration and trial type, seen in the right-hand side of Figure 3 and reflected by the model coefficient foil:semantic:adjacent. The source of this interaction can be seen in Figure 4, where the response rate to foil items is plotted against trial type and spatial configuration. A parallel analysis of the response to targets shows no statistically reliable difference between the phonetic-adjacent and the phonetic-separated conditions (not shown), indicating that the difference in d' between those conditions (seen in the right-hand side of Figure 3) is attributable to listener responses to foil items, consistent with the fact that the model coefficient for target:semantic:adjacent is not statistically significantly different from zero.

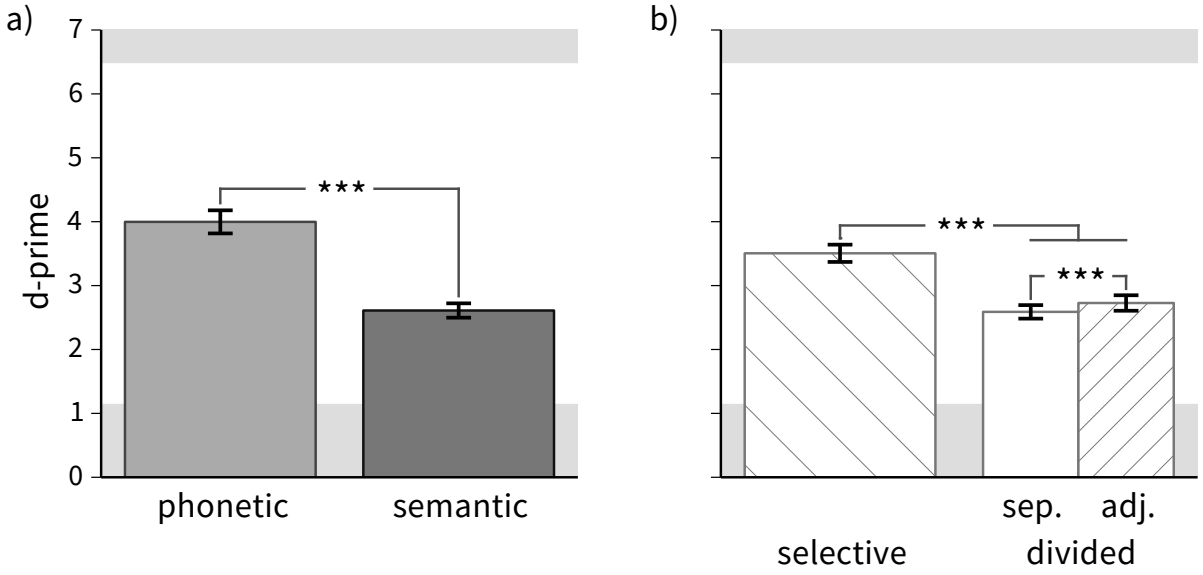

Figure 2: Barplots of mean listener sensitivity (d' scale) ± 1 standard error of the mean for the main effects "trial type" and "attentional configuration." Background shading indicates chance and ceiling performance levels. Brackets indicate the presence of corresponding coefficients in the statistical model that are significantly different from zero; * indicates p<0.05, ** indicates p<0.01, *** indicates p<0.001. (a) Main effect of trial type (phonetic versus semantic trials). The difference corresponds to the significant model coefficients target:semantic (indicating better target detection in phonetic trials) and foil:semantic (indicating higher tendency in the phonetic condition to incorrectly identify of foil items as targets). (b) Main effect of attentional configuration (selective attention to one stream versus divided attention to two streams). The selective versus divided difference (upper bracket) corresponds to the significant model coefficients target:selective (indicating better target detection in selective attention trials), and foil:selective (indicating higher tendency in divided attention trials to incorrectly respond to foil items). The adjacent versus non-ajacent difference (lower bracket) corresponds to the significant model coefficient foil:adjacent (indicating higher tendency in spatially non-adjacent divided attention trials to incorrectly respond to foil items).

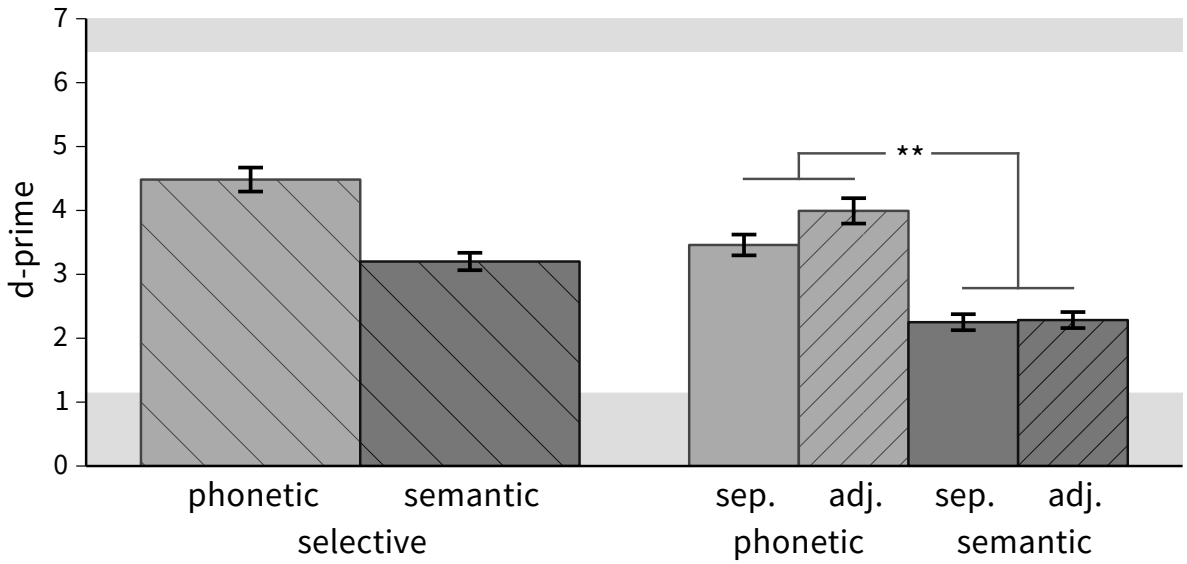

Figure 3: Barplots of mean listener sensitivity (d' scale) ± 1 standard error of the mean for the interactions between "trial type" and "attentional configuration." Background shading indicates chance and ceiling performance levels. Brackets indicate the presence of corresponding coefficients in the statistical model that are significantly different from zero; * indicates p<0.05, ** indicates p<0.01, *** indicates p<0.001. The bracket corresponds to the model coefficient foil:semantic:adjacent, which indicates increased responses to foils (i.e., more false alarms and thus lower d') in phonetic divided attention trials with spatially non-adjacent attended streams (light gray plain bar lower than light gray hatched bar, but dark gray bars equal).

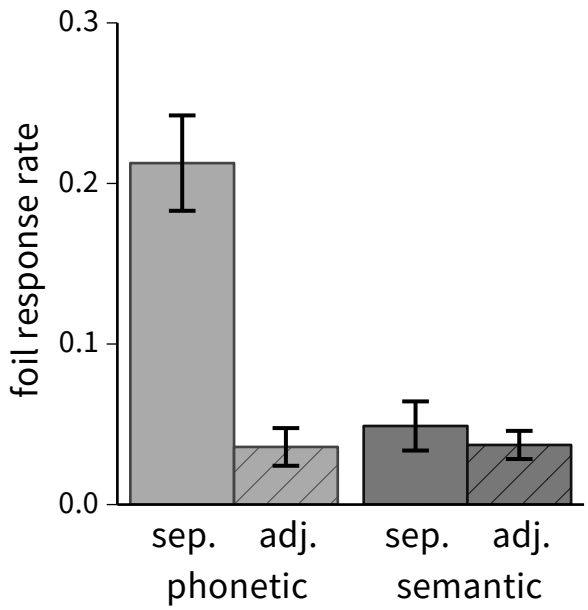

Figure 4: Barplots of mean listener response rate to foil words (oddballs occurring in to-be-ignored streams) ± 1 standard error of the mean for divided attention trials, showing the effect of attending spatially adjacent streams versus attending spatially separated streams (with one or more to-be-ignored streams interposed).

In addition to the question of listener sensitivity, there is a further question regarding the distribution of correct responses between the two attended streams in the divided attention trials. In other words, listeners might be able to achieve reasonably good d' scores by simply ignoring one of the two to-be-attended streams. The distribution of trials having at least one correct response to a target in *each* of the to-be-attended streams is shown in Table IV. The results show a much higher level of multi-stream detection in the phonetic trials than in the semantic trials, and a more modest difference between conditions where attended streams are spatially adjacent versus spatially separated. Nonetheless, there is some evidence of attention to both streams even in the most difficult conditions (semantic trials with non-adjacent attended streams).

Table IV: Distribution of trials showing evidence of attention to both of the to-be-attended streams (data pooled across subjects).

Experimental condition	Trials w/ hits in both attended streams
phonetic + adjacent	188 / 234 (80%)
phonetic + non-adjacent	159 / 234 (70%)
semantic + adjacent	139 / 468 (30%)
semantic + non-adjacent	123 / 468 (26%)

C. Discussion

This experiment shows a dramatic difference in listener oddball detection ability between trials where non-oddball stream items were repetitions of a single word per stream (phonetic trials) and trials where stream items were different words united by a semantic category relation (semantic trials). This result is unsurprising given the expectation that the phonetic condition should only require identification of a single deviant speech sound to identify oddballs, whereas the semantic condition should require matching the phonological information to a lexical item and making a category-membership judgment about that lexeme with reference to the category of the stream in which the word occurred. This difference should entail a difference in processing time and cognitive load that could certainly have a negative effect on performance in the semantic condition.

Another expected result was that responses to targets were more likely, and responses to foils slightly less likely, in selective attention trials compared to divided attention trials. This was predicted based on previous literature, as well as on the intuitive expectation that it is easier to monitor one stream for oddballs than to monitor two (although precisely why this is so is one of the questions at hand). Although listener performance in divided attention trials is well above chance even in the more difficult semantic condition, we hesitate to take this as evidence for truly parallel divided attention to different spatial locations, based on the large discrepancy between phonetic and semantic trial types in listener ability to respond to targets in multiple streams on the same trial, as seen in Table IV. We return to this point below.

Perhaps the most interesting result from this experiment is that responses to foils are more likely in trials with attention to non-adjacent streams, but only in phonetic trials (cf. the model coefficient foil:semantic:adjacent and Figure 4). This finding could be interpreted in a number of ways. First, if listeners are using a single auditory spotlight and rapidly switching between attended streams, the increased foil response could be indicative of a "sweeping across space" manner of attentional switching that leads to misallocation of foils in interposed streams to the flanking to-be-attended streams. Another interpretation is that listeners achieve divided attention by broadening their attentional spotlight, effectively encompassing a to-be-ignored stream that is interposed between two to-be-attended streams (cf. the findings of Hafter *et al.*, 2013, discussed in Section II., showing a trend toward better performance with smaller separation angles). A third possibility is that simultaneous auditory spotlights are deployed in parallel but are characterized by a spatial roll-off, and that when a to-be-ignored stream is sandwiched between two to-be-attended streams, the overlapping edges of the two attended stream spotlights cause the oddballs in the interposed to-be-ignored stream to be wrongly attributed to one of the flanking attended streams. However, all three of these interpretations are inconsistent with the lack of elevated foil responses in the semantic trials with spatially separated attended streams, so unless there is some other relevant factor present in the semantic trials but not in the phonetic trials that suppresses response to foils in spatially interposed to-be-ignored streams, we are left without a clear fit between our findings and any of the several models of auditory spatial attention.

One shortcoming of this experiment — and a possible cause of the aforementioned difference in responses to interposed foils — is that the task type (phonetic versus semantic) is confounded with a difference in the overall complexity of the auditory scene. In phonetic trials, a single base word was repeated in each stream, making oddballs *in any stream* more likely to stand out against the background of normal (non-oddball) items, potentially triggering exogenous reorientations of attention and giving rise to the observed increase in responses to foil items. In this light, the fact that foil responses in phonetic trials were higher when attended streams were

spatially separated might be explained as a pattern of exogenous reorientations triggered by highly salient foils, combined with spatial misallocation of the foils due to one of the mechanisms discussed above. Frequent exogenous reorientations in the phonetic trials but not in the semantic trials may also account for relatively high ratio of phonetic trials with hits in both attended streams (cf. Table IV). Seen this way, the results of Experiment 1 seem most consistent with a narrow attentional spotlight with rapid switching, since exogenous reorientation by definition involves a change in spatial locus of attention. Note that the arhythmicity of word onsets within each stream was designed to discourage temporal expectancy that would naturally lead to a rapid switching strategy, but listeners may have used such a strategy nonetheless, either because it is the only possible strategy or merely the most effective one. What is less clear is whether exogenous reorientation accounts for *all* cases of successful monitoring of two streams (even in the semantic trials), or merely accounts (in whole or in part) for the better performance in phonetic divided attention trials when compared to semantic ones.

V. EXPERIMENT 2

To overcome the confound between phonetic/semantic task type and overall acoustic scene complexity present in Experiment 1, in this experiment the phonetic condition was replaced with a second semantic condition, in which each category had only three members instead of six. There were still always twelve words per stream on each trial, so in the six-word condition each word occurred twice per trial (as in Experiment 1), and in the three-word condition each word occurred four times per trial (modulo replacement by oddballs). This still comprises a difference in overall complexity of the acoustic scene, but now both conditions are semantic tasks, and should require similar amounts and types of processing by the listener to carry out the task. Experiment 2 also introduced an additional manipulation of semantic "congruence": on some divided attention trials, the base words of the two attended streams had the same

category (though the order and timing of words was independent between the two streams).

A. Methods

Separate pre-experiment categorization tasks were provided for the small (3-word) categories and large (6-word) categories. All participants completed these tasks without errors, and passed the training procedure.

1. Participants

Seventeen participants (9 female) were recruited for this experiment, seven of whom had previously participated in Experiment 1; the delay between experiments was several months. One participant was excluded post-hoc based on extremely poor performance (hit rate of less than 0.5 in a selective attention control condition). Of the remaining sixteen participants, the age range was 18-31 (mean 25). Half the participants were presented the small-category trials first; the remainder were presented the large-category trials first.

2. Stimuli

Stimuli comprised sets of three or six monosyllabic words in each of eight categories (four 3-word categories and four 6-word categories; see Appendix for word lists). The categories did not statistically differ in the lexical frequency of their words, nor in the phonological neighborhood density of their words, nor the mean uniphone or biphone frequencies of their words. Statistical summaries of these comparisons are given in Table V. Unlike Experiment 1, where the extra semantic categories not in use during a trial block were used as a source of oddball words, in this experiment a dedicated set of 48 additional words were recorded for use as oddball items.

Trials were constructed by assigning each category to a spatial location; the assignment of category to location was held fixed within each experimental block, with the exception of

Table V: Summary of analysis of variance results for some lexical properties of the semantic categories in Experiment 2. Phonotactic probabilities were calculated using an online tool described by Vitevitch and Luce (2004); lexical frequency and neighborhood density data were drawn from Sommers (2014).

lexical property	**3-word categories**	**6-word categories**
lexical frequency	$F(3,8)=0.40$, $p=0.76$	$F(3,20)=1.03$, $p=0.40$
phonolgical neighborhood density	$F(3,8)=0.31$, $p=0.82$	$F(3,20)=0.20$, $p=0.90$
mean uniphone frequency	$F(3,8)=0.86$, $p=0.50$	$F(3,20)=0.22$, $p=0.89$
mean biphone frequency	$F(3,8)=1.59$, $p=0.27$	$F(3,20)=0.64$, $p=0.60$

divided attention trials in the "congruent" condition (in which case one of the category-location mappings was changed to the duplicated category while the others remained unchanged). Distribution of targets and foils was the same as in Experiment 1 (3–4 oddballs per trial, comprising 2–3 targets and 0–2 foils).

3. Procedure

As in Experiment 1, each trial's visual prompt indicated the correspondence between spatial location and stream category, and cued the participant to which streams were to-be-attended. There were six blocks of twenty trials each yielding a block length of about 5 minutes, and total of 120 trials.

4. Statisical analysis

Listener sensitivity was again modeled with generalized linear mixed-effects regression. The model predicted probability of listener button press for each word in the trial, based on "word type" (i.e., whether the word was a target, foil, or neither), whether the trial used three-word versus six-word categories ("size"), whether the attended stream categories were the same or different ("congruence"), and whether the attended streams were adjacent or spatially separated ("adjacency"). A random effect for participant was also included; the model equation is seen in

(5):

$$\Phi^{-1}(y_{ij}) = \beta_0 + \beta_1 W_{ti} + \beta_2 W_{fi} + \beta_3 Z_i + \beta_4 A_i + \beta_5 C_i + \ldots + S_{0j} + \epsilon_{ij} \quad (5)$$

In (5), Z_i is the indicator variable for "size" (denoted as three in Table VI), A_i is the indicator variable for "adjacency" (adjacent in Table VI), and C_i is the indicator variable for "congruence" (congruent in Table VI). All other terms are interpreted as in Equation (4). As in Experiment 1, the "word type" predictor was coded as a 3-level factor (target, foil, neither) with treatment coding and word.type=neither as baseline. The other three fixed-effects predictors ("size," "adjacency," and "congruence") were 2-level factors with deviation coding. Thus in Table VI three indicates the difference between three- and six-word trials (three minus six), adjacent indicates the difference between trials where attended streams were adjacent versus non-adjacent (adjacent minus non-adjacent), and congruent indicates the difference between trials where attended stream categories were congruent versus incongruent (congruent minus incongruent).

Although selective attention trials were included in Experiment 2, they were not analyzed as part of the statistical model reported here. This exclusion was done for two reasons: first, the selective attention trials were intended as a replication and control condition, to ensure that our results were comparable to those in Experiment 1. In fact, performance in selective versus divided conditions was quite similar to Experiment 1, so further analysis of the difference in performance between selective- and divided-attention conditions was deemed unnecessary. The second reason for excluding the selective attention trials is that the experimental manipulation "congruence" (sameness of semantic category of the attended streams) is conceptually meaningless when only one stream is attended. All other aspects of the modeling were identical to Experiment 1.

Using the approach to calculating d' described in Section III.D., perfect performance across all trials in this experiment yields a d' ceiling of 6.26; the highest performance of any subject in any condition was a d' of 5.26 in the selective attention three-word category trials. As mentioned in Section III.D., chance performance for Experiment 2 is conservatively estimated as

a d' value of 1.12. The lowest performance of any subject was a d' of 1.89, in the condition with six-word categories and divided attention to spatially separated streams, suggesting that none of the experimental conditions were so difficult that subjects had to resort to random response strategies.

B. Results

The model summary for listener responses is seen in Table VI, and corresponding barplots are shown in Figures 5 and 6. The model coefficients can be interpreted in similar fashion to the model for Experiment 1. Recall that a positive value for a coefficient containing target indicates *higher* sensitivity (increased hit rate), whereas a positive value for a coefficient containing foil indicates *lower* detection sensitivity (increased false alarm rate).

The variance estimated to account for differences between subjects was again quite small (standard deviation of 0.078 on a d' scale, compare 0.098 in the model for Experiment 1), suggesting that performance across subjects was again extremely consistent. Baseline response levels again show an expected pattern: sensitivity to targets was generally high (coefficient of 3.33, compare 3.31 from Experiment 1), response to foil items was lower (coefficient of 1.23; compare 0.97 from Experiment 1), and responses to non-target non-foil items were quite rare (coefficient of −2.69, compare −2.62 from Experiment 1). All baseline coefficients were significantly different from zero, though recall that these coefficients reflect treatment contrasts (not differences between conditions) so significant difference from zero is expected and unilluminating.

Among the coefficients for response bias, there is a small bias to respond less in the trials with three-word categories than in the trials with six-word categories (coefficient three is negative), and a slightly larger bias to respond less in trials in which attended streams are adjacent and have congruent categories (coefficient adjacent:congruent is negative). As in Experiment 1, the bias is attributable to differences in "stray" responses in those conditions: fewer stray responses in three-word trials (476) than six-word trials (612), and fewer stray responses in trials with

Table VI: Model summary predicting listener button presses in Experiment 2. * indicates p<0.05, ** indicates p<0.01, *** indicates p<0.001. † indicates significant coefficients (at the p<0.001 level) that are based on treatment coding; significance for these coefficients is expected and should be interpreted differently than the other coefficients (see text for explanation). SE = standard error of the coefficient estimate.

Indicator	Predictor name	Coef.	SE	z	p	Signif.
Baseline response levels						
	neither (Intercept)	−2.69	0.03	−96.15	<0.001	†
W_{ti}	target	3.33	0.03	115.31	<0.001	†
W_{fi}	foil	1.23	0.05	23.15	<0.001	†
Effect of manipulations on response bias						
Z_i	three	−0.11	0.04	−2.75	0.006	**
A_i	adjacent	−0.01	0.04	−0.14	0.892	
C_i	congruent	−0.03	0.04	−0.67	0.502	
$Z_i : A_i$	three:adjacent	0.05	0.08	0.67	0.505	
$Z_i : C_i$	three:congruent	−0.08	0.08	−1.05	0.292	
$A_i : C_i$	adjacent:congruent	−0.29	0.08	−3.70	<0.001	***
$Z_i : A_i : C_i$	three:adjct:congr	−0.10	0.16	−0.60	0.548	
Effect of manipulations on response to targets						
$W_{ti} : Z_i$	target:three	0.49	0.06	8.52	<0.001	***
$W_{ti} : A_i$	target:adjacent	0.09	0.06	1.49	0.136	
$W_{ti} : C_i$	target:congruent	0.07	0.06	1.26	0.206	
$W_{ti} : Z_i : A_i$	target:three:adjct	−0.19	0.11	−1.69	0.091	
$W_{ti} : Z_i : C_i$	target:three:congr	−0.10	0.11	−0.91	0.364	
$W_{ti} : A_i : C_i$	target:adjct:congr	0.95	0.11	8.23	<0.001	***
$W_{ti} : Z_i : A_i : C_i$	target:three:adjct:congr	0.21	0.23	0.92	0.359	
Effect of manipulations on response to foils						
$W_{fi} : Z_i$	foil:three	0.40	0.11	3.73	0.000	***
$W_{fi} : A_i$	foil:adjacent	−0.48	0.11	−4.57	<0.001	***
$W_{fi} : C_i$	foil:congruent	0.04	0.11	0.34	0.736	
$W_{fi} : Z_i : A_i$	foil:three:adjct	−0.13	0.21	−0.60	0.547	
$W_{fi} : Z_i : C_i$	foil:three:congr	0.11	0.21	0.51	0.612	
$W_{fi} : A_i : C_i$	foil:adjct:congr	−0.72	0.21	−3.43	0.001	***
$W_{fi} : Z_i : A_i : C_i$	foil:three:adjct:congr	−0.29	0.43	−0.69	0.489	

adjacent and congruent attended streams (98) compared to both trials with non-adjacent congruent streams (154) and trials with adjacent incongruent streams (160).

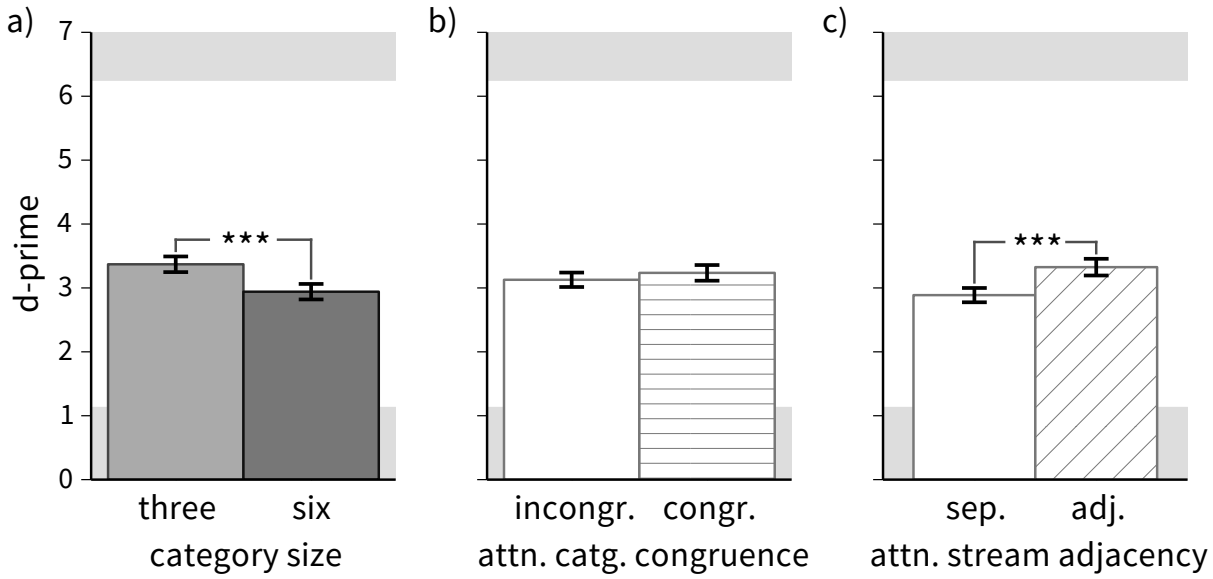

Figure 5: Barplots of mean listener sensitivity (d' scale) ± 1 standard error of the mean for the main effects "size," "congruence," and "adjacency" in Experiment 2. Background shading indicates chance and ceiling performance levels. Brackets indicate the presence of corresponding coefficients in the statistical model that are significantly different from zero; * indicates p<0.05, ** indicates p<0.01, *** indicates p<0.001. (a) Main effect of category size (three versus six words per category). The difference corresponds to the significant model coefficients target:three (positive, indicating better target detection in trials with three-word categories) and foil:three (positive, indicating higher tendency in the three-word condition to incorrectly identify of foil items as targets). (b) Main effect of attended category congruence. None of the relevant model coefficients (target:congruent and foil:congruent) are statistically significantly different from zero. (c) Main effect of spatial adjacency of the attended streams. The difference corresponds to the significant model coefficient foil:adjacent, indicating higher tendency to incorrectly identify foil items as targets when the attended streams are spatially separated.

The coefficient target:three indicates a significant main effect for category size, in the form of better target detection in trials with three-word categories (cf. Figure 5a). This effect is somewhat mitigated by the increased tendency to respond to foil items in trials with three-word categories, as indicated by the positive coefficient for foil:three. There is also a main effect for adjacency, driven by a decreased tendency to respond to foil items when attended streams are spatially adjacent (cf. the negative model coefficient for foil:adjacent, and Figure 5c).

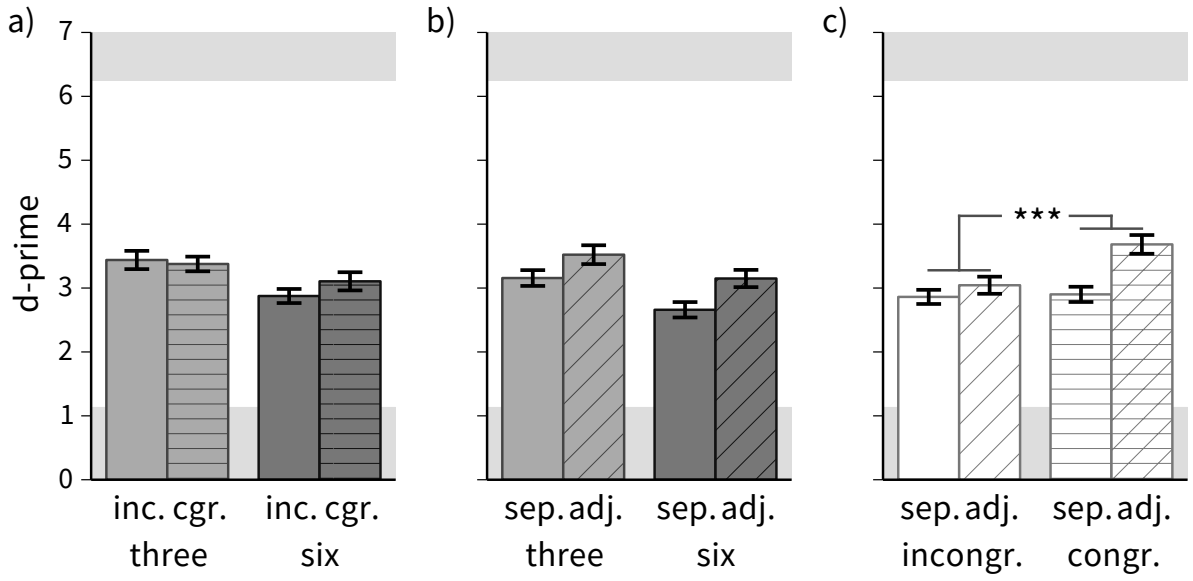

Figure 6: Barplots of mean listener sensitivity (d' scale) ± 1 standard error of the mean for the interactions among "size," "congruence," and "adjacency." Background shading indicates chance and ceiling performance levels. Brackets indicate the presence of corresponding coefficients in the statistical model that are significantly different from zero; * indicates p<0.05, ** indicates p<0.01, *** indicates p<0.001. (a) Plot of interaction between "size" (three- versus six-word categories) and "congruence" (same versus different categories in attended streams). None of the relevant model coefficients (target:three:congruent and foil:three:congruent) are statistically significantly different from zero. (b) Plot of interaction between "size" and "adjacency" (attended streams spatially separated or adjacent). None of the relevant model coefficients (target:three:adjacent and foil:three:adjacent) are statistically significantly different from zero. (c) Plot of interaction between "congruence" and "adjacency." The positive model coefficient target:adjacent:congruent and negative coefficient foil:adjacent:congruent indicate better target detection and fewer responses to foils when attended streams are both adjacent and congruent.

There is also an interaction between "congruence" and "adjacency" (cf. Figure 6c). When attended streams are both spatially adjacent and semantically congruent, responses to targets are more likely (coefficient target:adjacent:congruent is positive) and responses to foils are less likely (coefficient foil:adjacent:congruent is negative). Indeed, these two coefficients have the largest magnitude of any in the model (excluding baselines), suggesting the interaction is indeed a strong one.

The question of how often listeners had correct responses to targets in both of the to-be-attended streams is addressed in Table VII. Results are broadly similar to those seen in Experiment 1, in that fewer trials show target detection in both attended streams when the streams are spatially separated, and there is a difference between the 3-word and 6-word conditions parallel to the difference between phonetic and semantic trials in Experiment 1. There is also a trend toward higher detection of targets in both attended streams when the streams share the same category (values in "congruent" column generally higher than those in "incongruent" column).

Table VII: distribution of trials showing hits in both to-be-attended streams for Experiment 2 (data pooled across subjects).

Experimental condition	**Congruent**	**Incongruent**
3-word + adjacent	150 / 204 (74%)	123 / 204 (60%)
3-word + separated	119 / 204 (58%)	127 / 204 (62%)
6-word + adjacent	122 / 204 (60%)	96 / 204 (47%)
6-word + separated	101 / 204 (50%)	68 / 204 (33%)

C. Discussion

The finding that responses to both targets and foils were more likely in trials with 3-word categories than 6-word categories directly parallels the finding from Experiment 1 that responses to both targets and foils are more likely in the phonetic condition than the semantic condition. One explanation for this would be that the reduced complexity of the acoustic scene

in the 3-word condition led to an increase in the salience of oddballs (whether target or foil) in all streams, thereby drawing listener attention to the foil items and increasing the false alarm rate. However, the condition in which attended stream categories were congruent represents an intermediate level of acoustic scene complexity, since it involved four spatial streams but only three sets of words comprising those streams, reducing the number of non-oddball items from 12 or 24 in the 3-word and 6-word conditions (respectively) down to 9 and 18 items (respectively). According to the statistical model this difference does not affect the target or foil response rates (cf. coefficients for target:three:congruent and foil:three:congruent, and Figure 6a).

An alternative explanation for the finding that responses to both targets and foils were more likely in trials with 3-word categories than 6-word categories is that participants are using short-term memory to keep track of the pronunciation of the current attended stream's words, and there is a difference in the feasibility of this strategy between the 3-word and 6-word conditions. In other words, when the set of possible in-category words is smaller, listeners may be able to memorize all the words in the acoustic scene sufficiently well to detect oddballs on the basis of deviant sound patterns, without having to map the deviant sound patterns to a specific lexeme and compare that lexeme to the category associated with the spatial location from which it originated. The fact that category-location mappings were held constant within experimental blocks may have contributed to the feasibility of this strategy. However, although this explanation might account for a difference in target hit rate or in reaction time, it is unclear why responses to foils would also be higher in the 3-word condition (without further appeal to increased salience of foils, which implies a role for the overall complexity of the acoustic scene). When we also consider the high response rate to foil items in the 3-word condition in which attended streams are spatially separated (tallest bar in Figure 7b), combined with the differences between 3-word and 6-word conditions seen in Table VII, the conclusion that oddballs are more salient in the 3-word condition is even more difficult to resist, especially since this finding parallels the increase in response to foils in the phonetic-and-spatially-separated condition of

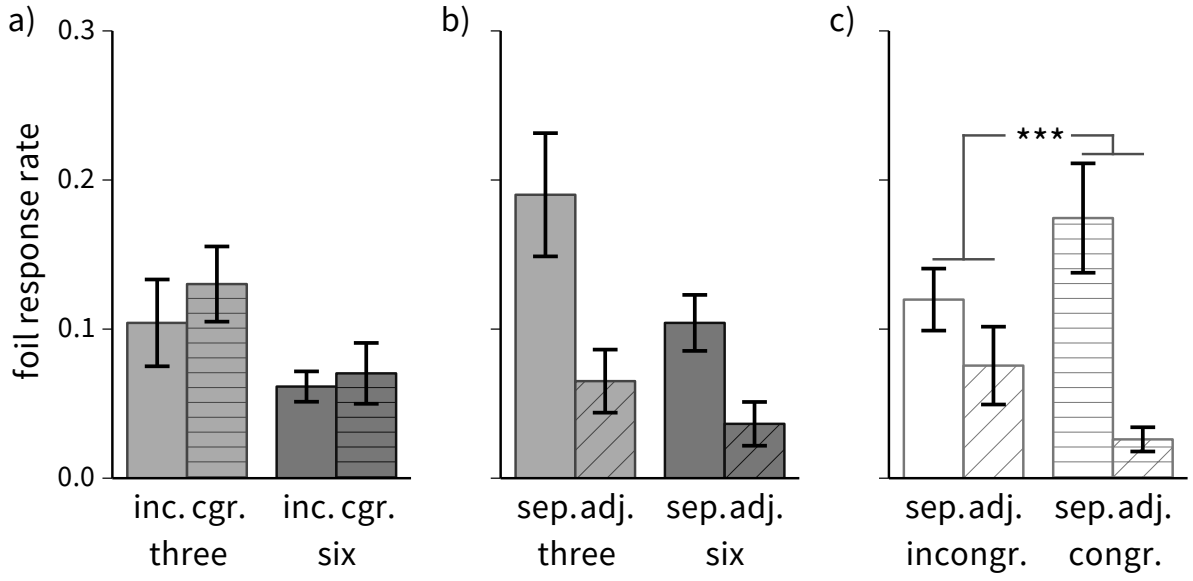

Figure 7: Barplots of mean listener response rate to foil words (oddballs occurring in to-be-ignored streams) ± 1 standard error of the mean for the two-way interactions among category size, attended stream adjacency, and attended stream category congruence.

Experiment 1. Taken together, these results give the impression that the reduced complexity of the acoustic scene in the 3-word condition allowed listeners to rely on phonetic information to accomplish a nominally semantic oddball detection task, possibly by the mechanism proposed to account for the Experiment 1 results (viz., reduced scene complexity leading to higher oddball salience, thus more exogenous reorientations to targets that would otherwise have been missed). Further experiments are needed to establish whether this proposed explanation accurately reflects listener strategy.

The finding that sensitivity is better (both in response to targets and suppression of response to foils) when attended streams are adjacent and attended stream categories are congruent suggests that, in such trials, listeners may be treating the adjacent streams as a single, diffuse spatial location or origin. Consistent with this view is the fact that, in cases where the attended streams have the same category but are spatially separated, the response to interposed foils is dramatically increased (much more so than the corresponding response to foils when the attended streams have different categories; cf. the two tallest bars of Figure 7c). In other words,

the semantic relationship between the to-be-attended streams seems to be triggering an attentional strategy that is non-ideal given the spatial configuration of the to-be-attended streams, and is suggestive of a perceptual architecture in which stream formation may be influenced by the linguistic content of the auditory scene, even to the point of overriding binaural cues that would normally suffice to segregate the sources into separate auditory objects.

Finally, the finding that responses to foils are less likely when attended streams are spatially adjacent was much weaker in the semantic condition of Experiment 1 than in this experiment (compare Figures 2b and 5c). However, this can be attributed to the inclusion of the "attended categories same/different" condition, which, through the interaction mentioned above, seems to be driving the apparent main effect of adjacency in Experiment 2 (cf. Figure 6c).

VI. POST-HOC ANALYSES

To further probe the hypothesis that listeners were leveraging phonetic information in the 3-word condition to accomplish a nominally semantic oddball detection task, we calculated reaction times for both experiments. The expectation was that, if listeners are able to (partially or completely) rely on phonetic information to accomplish the semantic task, and the ability to do so differed between the 3-word and 6-word conditions of Experiment 2, there should be a corresponding difference in reaction times between these two conditions. Specifically, the reaction times in the 3-word condition ought to be faster than those in the 6-word condition, but not as fast as the reaction times in the purely phonetic condition of Experiment 1.

A. Results

Barplots of the reaction times for Experiments 1 and 2 are shown in Figure 8. Only reaction times for "hit" responses were included; central tendency was calculated by taking the peak of a χ^2 distribution fitted to the reaction times for each subject in each condition (cf. Figure 9).

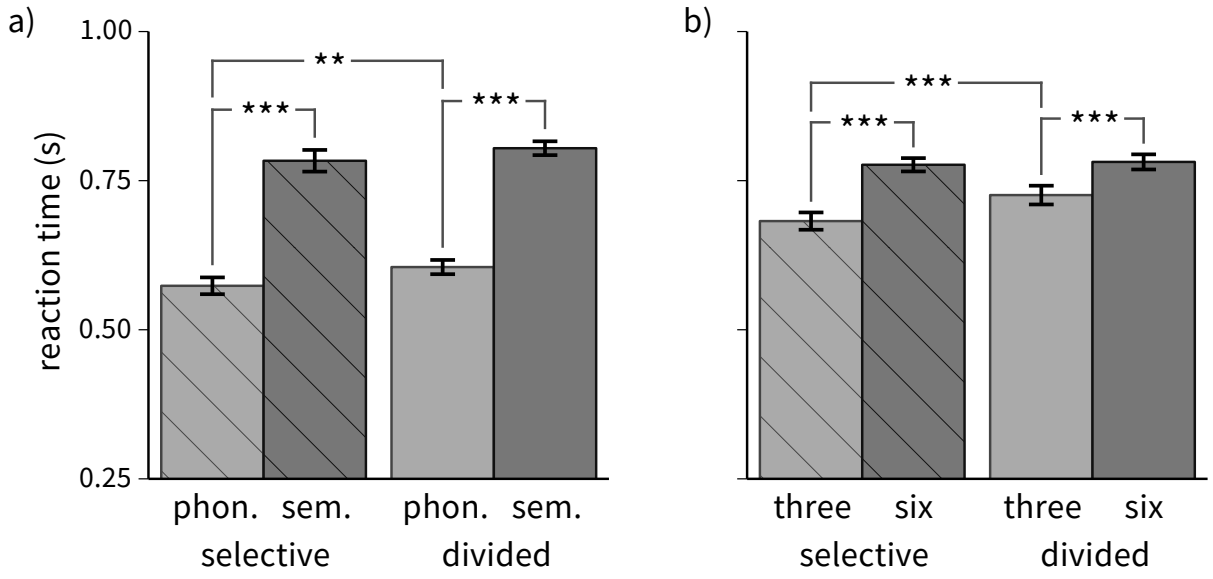

Figure 8: Barplots of listener reaction time ± 1 standard error of the estimate in Experiments 1 and 2. Brackets indicate significant differences (Bonferroni-corrected p-values from post-hoc pairwise t-tests): * indicates $p<0.05$, ** indicates $p<0.01$, *** indicates $p<0.001$. (a) Reaction times for the phonetic and semantic conditions of Experiment 1, separated by number of attended streams ("selective" = attend one stream; "divided" = attend two streams). (b) Reaction times for the 3-word and 6-word conditions of Experiment 2, separated by number of attended streams. As predicted, the reaction times for the 3-word condition are significantly shorter than for the 6-word condition, and the value for the 3-word condition falls between the values for the phonetic and semantic conditions of Experiment 1.

Results are consistent with the stated prediction: reaction times for the 3-word condition in Experiment 2 were statistically significantly shorter than reaction times for the 6-word condition. Moreover, reaction times for the 3-word condition fell in between the reaction times for the phonetic and semantic conditions in Experiment 1. Because some, but not all, subjects from Experiment 1 participated in Experiment 2, neither a standard independent samples t-test nor a paired samples t-test is possible using all the data. However, when comparing reaction times using a paired-samples t-test for only the six subjects who participated in both experiments, we indeed see significant differences between the 3-word condition of Experiment 2 and the phonetic condition of Experiment 1 (p=0.032 for the selective attention trials, and p=0.006 for the divided attention trials; this comparison not shown in Figure 8). Reaction times in the 6-word condition of Experiment 2 do not significantly differ from the semantic condition in Experiment 1 (p=0.150 for the selective attention trials, and p=0.280 for the divided attention trials; this comparison not shown in Figure 8).

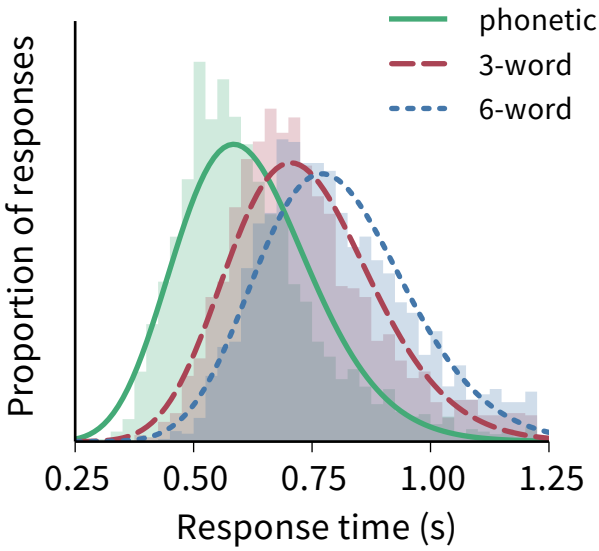

Figure 9: (Color online) Normalized histograms and χ^2 fits of reaction times in the phonetic condition of Experiment 1 and the 3-word and 6-word conditions of Experiment 2 (data pooled across subjects). The reaction times of the 3-word condition are intermediate between the phonetic and 6-word conditions, suggesting that participants may leverage phonetic information when possible to speed their responses in nominally semantic tasks.

B. Discussion

The reaction time results lend further support to the idea that listeners were leveraging phonetic information to make oddball judgments in the 3-word task of Experiment 2. However, it is unclear whether, in the 3-word condition of Experiment 2, listeners were relying on a mix of phonetic and semantic judgments to detect oddballs, or were taking an exclusively phonetic approach but were slowed down (relative to the phonetic condition in Experiment 1) by the increased complexity of the task. Of course, it is also possible that listeners were, on some trials, making oddball judgments purely on the basis of phonetic information even in some of the 6-word trials of Experiment 2 (or the semantic trials of Experiment 1). For example, if a target word occurred that began with [s] and by some chance none of the base category words in a particular category began with [s], a listener might make a relatively faster oddball judgment for that particular target. Such opportunities ought to be more common when there are fewer base words in each category, as there will be fewer unique phone sequences forming the acoustic backdrop against which the oddballs must be detected. In this light, it is easier to understand why listener performance and response time in the 3-word condition is intermediate between the phonetic condition of Experiment 1 (effectively a 1-word-per-stream condition) and the 6-word condition of Experiment 2, in spite of the fact that listeners were supposedly performing a semantic task in the 3-word condition.

VII. GENERAL DISCUSSION

The experimental paradigm used in these experiments allows the investigation of selective and divided auditory attention in a multi-stream environment. Although these experiments only involved four auditory streams at four spatial locations, the paradigm has been used successfully in studies of selective attention with up to twelve streams in a condition similar to the phonetic condition of Experiment 1, but using alphabet letters instead of ordinary words (Maddox *et al.*, 2012). That study also showed that listener performance was primarily

determined by informational masking, and the effect of energetic masking due to temporal overlap of tokens in different streams was negligible even at a presentation rate of 12 Hz (i.e., with a mean token length of around 430 ms, a mid-trial token would typically exhibit temporal overlap with five preceding and five following tokens — much more overlap than seen in the present study). In this study, the SNR of the attended words was relatively high (mean +3 dB in the better ear, not considering any additional release from masking resulting from the wide separation angles of the streams), and the lexical properties of the categories were carefully balanced (including elimination of polysemous words and minimal pairs). This means that listener inability to identify an oddball as an oddball is almost certainly attributable to limitations of the listener's attention, rather than inability to understand the word clearly, or mistaking the oddball word for a different, more familiar word that conformed (phonetically or semantically) to the stream category. In light of this, it is tempting to interpret the results of the present experiments — in particular the result that performance in the 3-word semantic condition was intermediate between the 1-word phonetic and the 6-word semantic conditions — as a case of differences in performance due to differences in informational masking arising from target-masker spectrotemporal similarity, where the masker is the set of background (non-oddball) words across all streams (cf. Kidd *et al.*, 2002; Durlach *et al.*, 2003; Iyer *et al.*, 2010; Calandruccio *et al.*, 2013, *inter alia*).

In addition, it is noteworthy to learn that a listener's ability to divide attention — particularly the ability to ignore irrelevant interposed streams — is so strongly affected by a linguistic property such as semantic coherence between the attended streams. To use an analogy, when seated across the table from three friends (let's call them Larry, Moe, and Curly) who all happen to be talking at once, these results suggest that Moe (the middle one) is harder to ignore if Larry and Curly both happen to be discussing baseball (i.e., their speech streams are semantically coherent) than if Larry were discussing the circus and Curly were discussing tai chi. Of course, words in natural speech do not have the same staggering of onsets across streams as seen in this paradigm, and words occurring later in natural sentences are often highly predictable from the

preceding context (unlike this paradigm, where all but the oddball words are known in advance). Nonetheless, the semantic coherence in this experimental paradigm at least somewhat approximates the coherent semantic relationships among the words in a natural sentence. Another interesting aspect of this paradigm is that participant responses to foils can be as informative as responses to targets. Indeed, the elevated responses to foils in ignored streams interposed spatially between attended streams in both the phonetic condition of Experiment 1 and the 3-word condition of Experiment 2 both inform the debate regarding models of auditory spatial divided attention, and seem to be most consistent with a "rapid switching" model. However, it remains a possibility that listeners use different strategies in different listening situations, so these results should not be treated as conclusive evidence against the "broadened spotlight" or "multiple parallel spotlights" models (indeed, the effect of semantic coherence of attended streams on the ability to ignore interposed streams would seem to favor a broadened spotlight model). Moreover, the proper interpretation of false alarm responses is itself an open question: here we have implicitly been interpreting responses to foil items as detection of a deviant combined with irresistible exogenously-driven response to the deviant, but foil responses could conceivably also be interpreted as detection of deviants combined with localization error. We believe this latter interpretation to be unlikely given the large separation angles used in these experiments, but we cannot rule it out on the grounds of these experiments alone. Finally, even though we believe the responses to foils (and quite possibly many of the responses to targets) are a reflection of exogenously-driven attentional shifts, it is important to note that the lack of response to foils must not be interpreted as the absence of such exogenous reorientations. This is because a foil item may draw a participant's attention even if the ensuing behavioral response to it is suppressed. This suggests the need for supplementary methodologies (such as imaging studies tracking neural correlates of attention switches) to better address questions of exogenous reorientations of attention.

The fundamental question of this study was whether attending to different linguistic aspects of a speech signal would impact a listener's ability to direct their attention spatially and to divide

their attention between multiple spatially-defined streams. This question remains to some degree unanswered. Differences in task difficulty may have led to subtle differences in spatial release from masking among the various conditions, which could have impacted performance. Moreover, in light of listeners' apparent ability to rely on phonetic information to speed their responses in nominally semantic tasks, there remains some question as to how different their attentional states were when performing in the phonetic and semantic conditions. Future experimental designs may overcome some of these challenges, as would studies of listeners' neural activity during such tasks.

In addition, differences in acoustic scene complexity emerged as an important consideration, possibly due to differences in oddball salience and the potential role of exogenous reorientations in responding to such "would-have-been-missed" oddballs. As such, we cannot definitively attribute performance differences in our phonetic and semantic tasks to differences in listener attentional state, without first accounting for the differences in the acoustic scene. Nonetheless, we did find evidence of divided attention (in the form of correct responses to targets in both to-be-attended streams in the same trial), and differences in this ability between experimental conditions (cf. Tables IV and VII). Whether this reflects "true simultaneous attention" to multiple locations is unclear: the rapid-button-press response paradigm should have minimized the extent to which listeners could rely on short-term memory to improve performance, but at the same time the possibility of attention to one location combined with exogenous reorientations is a plausible alternative account. In this regard our findings did not provide conclusive evidence for which model of divided attention is the correct one — if anything, our results suggested that both broadened spotlight and rapid switching strategies are possible, and their deployment may be determined by properties the task (e.g., congruence of the semantic categories of the attended streams). Further questions include whether listeners have a conscious choice in which strategy they deploy, and how to discriminate between two types of attention switches that might both be called "rapid switching": namely, endogenously versus exogenously triggered reorientations.

VIII. ACKNOWLEDGMENTS

This research was supported by NIH grants R01-DC013260 to Adrian KC Lee and T32-DC000033 to the Department of Speech and Hearing Sciences, University of Washington. The authors are grateful to two anonymous reviewers and the members of *[LABS]*^{N} for helpful suggestions on earlier drafts of this paper.

APPENDIX: WORD AND CATEGORY LISTS

Table VIII: Semantic categories and category items used in Experiment 1.

Animals	**Body**	**Clothes**	**Food & drink**	**Furniture**	**Plants**	**Weather**
bird	arm	belt	beer	bed	bark	breeze
cat	chin	dress	bread	chair	grass	cloud
cow	foot	hat	meat	couch	leaf	fog
mouse	leg	scarf	rice	desk	root	rain
pig	mouth	shirt	soup	lamp	stick	sky
snake	nose	suit	wine	rug	tree	wind

Table IX: Words used in the phonetic condition of Experiment 1.

branch	knee
cake	moss
duck	pants
fish	sink
fruit	stove
hail	wrist

Table X: Three-word categories used in Experiment 2.

Fruit	**Birds**	**Fish**	**Drinks**
lime	hawk	eel	wine
fig	duck	bass	juice
grape	goose	cod	tea

Table XI: Six-word categories used in Experiment 2.

Food	**Furniture**	**Weather**	**Colors**
bread	bed	hail	blue
stew	desk	rain	gray
meat	chair	frost	green
cake	lamp	wind	pink
fruit	stool	storm	red
rice	couch	cloud	tan

Table XII: Oddball words used in Experiment 2.

bear	arm	belt	bark
cat	chest	dress	branch
cow	chin	glove	bud
dog	ear	hat	seed
fox	foot	pants	leaf
goat	knee	purse	root
horse	leg	scarf	stem
mouse	mouth	shirt	stick
pig	nose	shoe	thorn
rat	teeth	skirt	tree
sheep	thigh	sock	trunk
snake	wrist	suit	vine

REFERENCES

Allport, D. A., Antonis, B., and Reynolds, P. (**1972**), "On the division of attention: A disproof of the single channel hypothesis," Q. J. Exp. Psychol. **24**(2), 225–235.

Bates, D., Maechler, M., and Bolker, B. (**2014**), "lme4: Linear mixed-effects models using S4 classes," version 1.1-7, URL http://CRAN.R-project.org/package=lme4 (date last viewed 2014-07-19).

Best, V., Gallun, F. J., Ihlefeld, A., and Shinn-Cunningham, B. G. (**2006**), "The influence of spatial separation on divided listening," J. Acoust. Soc. Am. **120**(3), 1506–1516.

Boersma, P. and Weenink, D. (**2014**), "Praat: Doing phonetics by computer," version 5.3.69, URL http://www.praat.org/ (date last viewed 2014-03-28).

Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (**2000**), "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**(2), 1065–1066.

Broadbent, D. E. (**1958**), *Perception and communication* (Pergamon Press, London).

Calandruccio, L., Brouwer, S., Van Engen, K. J., Dhar, S., and Bradlow, A. R. (**2013**), "Masking release due to linguistic and phonetic dissimilarity between the target and masker speech," Am. J. Audiol. **22**(1), 157–164.

Cherry, E. C. (**1953**), "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**(5), 975–979.

DeCarlo, L. T. (**1998**), "Signal detection theory and generalized linear models," Psychol. Methods **3**(2), 186–205.

Dupoux, E., Kouider, S., and Mehler, J. (**2003**), "Lexical access without attention? Explorations using dichotic priming," J. Exp. Psychol. Human **29**(1), 172–184.

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., and Kidd, G., Jr. (**2003**), "Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity," J. Acoust. Soc. Am. **114**(1), 368–379.

Eich, E. (**1984**), "Memory for unattended events: Remembering with and without awareness," Mem. Cognition **12**(2), 105–111.

Eriksen, C. W. and Hoffman, J. E. (**1972**), "Temporal and spatial characteristics of selective encoding from visual displays," Percept. Psychophys. **12**(2), 201–204.

Eriksen, C. W. and St. James, J. D. (**1986**), "Visual attention within and around the field of focal attention: A zoom lens model," Percept. Psychophys. **40**(4), 225–240.

Gallun, F. J., Mason, C. R., and Kidd, G., Jr. (**2007**), "Task-dependent costs in processing two simultaneous auditory stimuli," Percept. Psychophys. **69**(5), 757–771.

Glucksberg, S. and Cowen, G. N., Jr. (**1970**), "Memory for nonattended auditory material," Cognitive Psychol. **1**(2), 149–156.

Hafter, E. R., Xia, J., Kalluri, S., Poggesi, R., Hansen, C., and Whiteford, K. (**2013**), "Attentional switching when listeners respond to semantic meaning expressed by multiple talkers," Proc. Meet. Acoust. **19**, 050077.

Hickok, G. and Poeppel, D. (**2004**), "Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language," Cognition **92**(1-2), 67–99.

Hickok, G. and Poeppel, D. (**2007**), "The cortical organization of speech processing," Nat. Rev. Neurosci. **8**(5), 393–402.

Ihlefeld, A. and Shinn-Cunningham, B. (**2008**), "Spatial release from energetic and informational masking in a divided speech identification task," J. Acoust. Soc. Am. **123**(6), 4380–4392.

Iyer, N., Brungart, D. S., and Simpson, B. D. (**2010**), "Effects of target-masker contextual similarity on the multimasker penalty in a three-talker diotic listening task," J. Acoust. Soc. Am. **128**(5), 2998–3010.

Kidd, G., Jr., Mason, C. R., and Arbogast, T. L. (**2002**), "Similarity, uncertainty, and masking in the identification of nonspeech auditory patterns," J. Acoust. Soc. Am. **111**(3), 1367–1376.

Larson, E. D., McCloy, D. R., Maddox, R. K., and Pospisil, D. (**2014**), "expyfun: Python experimental paradigm functions," version 2.0.0, URL https://github.com/LABSN/expyfun (date last viewed 2014-04-09).

Lawrence, M. A. and Klein, R. M. (**2013**), "Isolating exogenous and endogenous modes of temporal attention," J. Exp. Psychol. Gen. **142**(2), 560–572.

Lawson, E. A. (**1966**), "Decisions concerning the rejected channel," Q. J. Exp. Psychol. **18**(3), 260–265.

Macmillan, N. A. and Creelman, C. D. (**2005**), *Detection theory: a user's guide* (Lawrence Erlbaum Associates, Mahwah, NJ), 2 ed.

Maddox, R. K., Cheung, W., and Lee, A. K. C. (**2012**), "Selective attention in an overcrowded auditory scene: Implications for auditory-based brain-computer interface design," J. Acoust. Soc. Am. **132**(5), EL385–EL390.

Moray, N. (**1959**), "Attention in dichotic listening: Affective cues and the influence of instructions," Q. J. Exp. Psychol. **11**(1), 56–60.

Moulines, É. and Charpentier, F. J. (**1990**), "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun. **9**(5-6), 453–467.

Norman, D. A. (**1969**), "Memory while shadowing," Q. J. Exp. Psychol. **21**(1), 85–93.

R Development Core Team (**2014**), "R: A language and environment for statistical computing," version 3.1.1, URL http://www.R-project.org/ (date last viewed 2014-07-10).

Rivenez, M., Darwin, C. J., and Guillaume, A. (**2006**), "Processing unattended speech," J. Acoust. Soc. Am. **119**(6), 4027–4040.

Sheu, C.-F., Lee, Y.-S., and Shih, P.-Y. (**2008**), "Analyzing recognition performance with sparse data," Behav. Res. Meth. **40**(3), 722–727.

Shinn-Cunningham, B. G., Kopco, N., and Martin, T. J. (**2005**), "Localizing nearby sound sources in a classroom: Binaural room impulse responses," J. Acoust. Soc. Am. **117**(5), 3100–3115.

Sommers, M. S. (**2014**), "WU Speech & Hearing Lab Neighborhood Database," URL http://neighborhoodsearch.wustl.edu/Home.asp (date last viewed 2014-03-21).

Treisman, A. M. (**1960**), "Contextual cues in selective listening," Q. J. Exp. Psychol. **12**(4), 242–248.

Vitevitch, M. S. and Luce, P. A. (**2004**), "A Web-based interface to calculate phonotactic probability for words and nonwords in English," Behav. Res. Meth. Ins. C. **36**(3), 481–487.

Wood, N. and Cowan, N. (**1995**a), "The cocktail party phenomenon revisited: How frequent are attention shifts to one's name in an irrelevant auditory channel?" J. Exp. Psychol. Learn. **21**(1), 255–260.

Wood, N. L. and Cowan, N. (**1995**b), "The cocktail party phenomenon revisited: Attention and memory in the classic selective listening procedure of Cherry (1953)." J. Exp. Psychol. Gen. **124**(3), 243–262.

Wood, N. L., Stadler, M. A., and Cowan, N. (**1997**), "Is there implicit memory without attention? A reexamination of task demands in Eich's (1984) procedure," Mem. Cognition **25**(6), 772–779.

LIST OF FIGURES

3 Barplots of mean listener sensitivity (d' scale) ± 1 standard error of the mean for the interactions between "trial type" and "attentional configuration." Background shading indicates chance and ceiling performance levels. Brackets indicate the presence of corresponding coefficients in the statistical model that are significantly different from zero; * indicates p<0.05, ** indicates p<0.01, *** indicates p<0.001. The bracket corresponds to the model coefficient foil:semantic:adjacent, which indicates increased responses to foils (i.e., more false alarms and thus lower d') in phonetic divided attention trials with spatially non-adjacent attended streams (light gray plain bar lower than light gray hatched bar, but dark gray bars equal).

4 Barplots of mean listener response rate to foil words (oddballs occurring in to-be-ignored streams) ± 1 standard error of the mean for divided attention trials, showing the effect of attending spatially adjacent streams versus attending spatially separated streams (with one or more to-be-ignored streams interposed).

5 Barplots of mean listener sensitivity (d' scale) ± 1 standard error of the mean for the main effects "size," "congruence," and "adjacency" in Experiment 2. Background shading indicates chance and ceiling performance levels. Brackets indicate the presence of corresponding coefficients in the statistical model that are significantly different from zero; * indicates p<0.05, ** indicates p<0.01, *** indicates p<0.001. (a) Main effect of category size (three versus six words per category). The difference corresponds to the significant model coefficients target:three (positive, indicating better target detection in trials with three-word categories) and foil:three (positive, indicating higher tendency in the three-word condition to incorrectly identify of foil items as targets). (b) Main effect of attended category congruence. None of the relevant model coefficients (target:congruent and foil:congruent) are statistically significantly different from zero. (c) Main effect of spatial adjacency of the attended streams. The difference corresponds to the significant model coefficient foil:adjacent, indicating higher tendency to incorrectly identify foil items as targets when the attended streams are spatially separated.

6 Barplots of mean listener sensitivity (d' scale) ± 1 standard error of the mean for the interactions among "size," "congruence," and "adjacency." Background shading indicates chance and ceiling performance levels. Brackets indicate the presence of corresponding coefficients in the statistical model that are significantly different from zero; * indicates p<0.05, ** indicates p<0.01, *** indicates p<0.001. (a) Plot of interaction between "size" (three- versus six-word categories) and "congruence" (same versus different categories in attended streams). None of the relevant model coefficients (target:three:congruent and foil:three:congruent) are statistically significantly different from zero. (b) Plot of interaction between "size" and "adjacency" (attended streams spatially separated or adjacent). None of the relevant model coefficients (target:three:adjacent and foil:three:adjacent) are statistically significantly different from zero. (c) Plot of interaction between "congruence" and "adjacency." The positive model coefficient target:adjacent:congruent and negative coefficient foil:adjacent:congruent indicate better target detection and fewer responses to foils when attended streams are both adjacent and congruent.

7 Barplots of mean listener response rate to foil words (oddballs occurring in to-be-ignored streams) ± 1 standard error of the mean for the two-way interactions among category size, attended stream adjacency, and attended stream category congruence.

8 Barplots of listener reaction time ± 1 standard error of the estimate in Experiments 1 and 2. Brackets indicate significant differences (Bonferroni-corrected p-values from post-hoc pairwise t-tests): * indicates $p<0.05$, ** indicates $p<0.01$, *** indicates $p<0.001$. (a) Reaction times for the phonetic and semantic conditions of Experiment 1, separated by number of attended streams ("selective" = attend one stream; "divided" = attend two streams). (b) Reaction times for the 3-word and 6-word conditions of Experiment 2, separated by number of attended streams. As predicted, the reaction times for the 3-word condition are significantly shorter than for the 6-word condition, and the value for the 3-word condition falls between the values for the phonetic and semantic conditions of Experiment 1.

9 (Color online) Normalized histograms and χ^2 fits of reaction times in the phonetic condition of Experiment 1 and the 3-word and 6-word conditions of Experiment 2 (data pooled across subjects). The reaction times of the 3-word condition are intermediate between the phonetic and 6-word conditions, suggesting that participants may leverage phonetic information when possible to speed their responses in nominally semantic tasks.