Published in final edited form as:
J Mem Lang. 2014 February 1; 71(1): 145–163. doi:10.1016/j.jml.2013.11.002.

Immediate effects of anticipatory coarticulation in spoken-word recognition

Anne Pier Salverda[*], **Dave Kleinschmidt**, and **Michael K. Tanenhaus**
Department of Brain and Cognitive Sciences, University of Rochester, United States

Abstract

Two visual-world experiments examined listeners' use of pre word-onset anticipatory coarticulation in spoken-word recognition. Experiment 1 established the shortest lag with which information in the speech signal influences eye-movement control, using stimuli such as "*The … ladder is the target*". With a neutral token of the definite article preceding the target word, saccades to the referent were not more likely than saccades to an unrelated distractor until 200–240 ms after the onset of the target word. In Experiment 2, utterances contained definite articles which contained natural anticipatory coarticulation pertaining to the onset of the target word (" *The ladder … is the target*"). A simple Gaussian classifier was able to predict the initial sound of the upcoming target word from formant information from the first few pitch periods of the article's vowel. With these stimuli, effects of speech on eye-movement control began about 70 ms earlier than in Experiment 1, suggesting rapid use of anticipatory coarticulation. The results are interpreted as support for "data explanation" approaches to spoken-word recognition. Methodological implications for visual-world studies are also discussed.

Keywords

Spoken-word recognition; Speech perception; Coarticulation; Visual-world paradigm

Introduction

At any point in time during the production of speech, a speaker's articulatory gestures are not only determined by the sound that is currently being articulated but also by preceding and upcoming sounds. This core property of speech is generally referred to as coarticulation. If the production of each speech sound was programmed and executed independently, articulation would proceed substantially more slowly, and the speed and efficiency of speech production would be compromised. Coarticulatory effects can be *perseverative*, when the production of a segment is affected by the production of a preceding segment, or *anticipatory*, when the production of a segment is affected by an upcoming segment. Both types of coarticulation affect the resulting acoustic signal.

In the present study, we examine the role of anticipatory coarticulation in spoken-word recognition. In particular, we examine whether and to what extent listeners can use coarticulatory information in the vowel of the definite article, "the", to anticipate the soundform of an upcoming word. For example, can listeners use information early in the vowel to generate predictions that the following word will begin with a particular sound (for

[*]Corresponding author. Address: Department of Brain and Cognitive Sciences, University of Rochester, Meliora Hall, Box 270268, Rochester, NY 14627-0268, United States. Fax: +1 585 442 9216. asalverda@bcs.rochester.edu (A.P. Salverda).

instance the fricative /f/ when the following word is “fish”) before the onset of the frication noise? If so, how quickly do listeners use such information? We focus on pre word-onset anticipatory coarticulation for empirical, theoretical and methodological reasons. We briefly consider each of these in turn before providing the relevant motivation and background information for our experiments.

Empirical, theoretical and methodological motivations

Anticipatory coarticulation, including coarticulation across word boundaries, is well-documented in the phonetics literature. However, there is relatively little empirical evidence about real-time processing and use of anticipatory coarticulation that precedes word onset. As we review later, most of the existing evidence focuses on a particular phonological environment, coronal place assimilation. Moreover, most work on coarticulation in spoken-word recognition has provided limited insight into the time course of these effects. The current work makes an empirical contribution by extending research on pre-onset coarticulation to a more general environment and by providing detailed information about time course.

How listeners make use of the acoustic correlates of pre-onset anticipatory coarticulation is of theoretical interest because it has the potential to provide evidence that could distinguish between the three major classes of models of spoken-word recognition: (1) pattern-recognition models with abstract representations; (2) exemplar-based pattern recognition models; and (3) data-explanation models that evaluate hypotheses about the state of the world according to how well those hypotheses predict the observed properties of the perceptual input. We return to this point in the general discussion. For now, it will suffice to foreshadow the form of the argument.

The first class of model, *pattern-recognition models* such as TRACE (McClelland & Elman, 1986) or Shortlist (Norris, 1994; Norris & McQueen, 2008) could use fine-grained phonetic detail, including that due to anticipatory coarticulation, to activate and help select among partially activated lexical candidates, thereby modulating competitor effects (e.g., competition from lexical neighbors) by providing information about an upcoming sound or sounds. Pattern-recognition models most naturally incorporate pre-onset coarticulation by using context sensitive features or other methods of enriching the phonetic detail captured in the front end of the model, for instance, acoustic templates based on multi-phoneme sequences.

In *exemplar models* (e.g., Goldinger, 1998; Johnson, 1997; Pierrehumbert, 2002), stored representations of all of the instances of words that listeners have encountered naturally encode fine-grained phonetic detail, including speaker-specific information and within-word coarticulation. Pre word-onset anticipatory coarticulation can be accounted for by assuming that exemplars can correspond to multi-word chunks. For instance, in addition to exemplars for the word “ladder”, an exemplar model might store exemplars for sequences of words with relatively high co-occurrence rates, such as “the ladder” and “a ladder”.

Data explanation models seek to find explanations for otherwise unexplained data, in particular “error signals” that arise from differences between the expected input (which is predicted by a generative model) and the observed input. For example, acoustic–phonetic properties of the vowel of the spoken word “the” that cannot be explained by predictions based on factors such as the speaker and the speech rate, could be explained by hypotheses about the upcoming word or sound that might account for the observed data.

We will argue that although use of pre-onset acoustic/phonetic cues is not incompatible with pattern-recognition and exemplar models, data-explanation approaches predict these effects

and provide the most natural account. Moreover, if there are pre-onset effects, coarticulation could serve as an empirical domain for exploring novel predictions that arise from data-explanation models.

The methodological motivation for the current work is that many inferences about the time course of language processing in the visual-world paradigm (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) depend upon assumptions about the minimal lag between the processing of acoustic information in the speech signal and the effect of that information on eye movements. This lag is used to define time windows within which to examine fixations that could be sensitive to events of interest in the speech signal.

Beginning with Tanenhaus et al. (1995), studies using the visual-world paradigm have typically assumed that it takes about 200 ms for information in the speech signal to influence eye movements to potential referents. The 200-ms estimate was initially motivated by findings from the vision literature demonstrating that it takes approximately 150–200 ms to program and execute a saccade in a display with more than two possible target locations (Hallett, 1986; Matin, Shao, & Boff, 1993). Because linguistically-mediated saccades would presumably be somewhat slower than the fastest visually-guided saccades, a 200-ms delay seemed an appropriately conservative estimate.

Assumptions about the lag between speech and eye movements often affect the interpretation of results from visual-world studies. For example, a line of research beginning with Sedivy, Tanenhaus, Chambers, and Carlson (1999) argues that listeners construct contrast sets on the basis of information available in the visual context and a pre-nominal scalar adjective *prior* to encountering the head noun (e.g., upon hearing the word “tall” in the instruction “Pick up the tall glass”). If effects of speech on eye-movement control[1] can occur with a lag that is considerably shorter than 200 ms, for instance within 100 ms as suggested by Altmann (2011), then the conclusions from studies like Sedivy et al. and many others would need to be re-evaluated. (For a recent example of a study on low-level speech perception in which the assumption of a 200-ms lag is essential for the interpretation of the data, see Mitterer & Reinisch, 2013, and discussion therein.)

Anticipatory coarticulation in spoken-word recognition

Coarticulation has been well documented in the articulatory and acoustic phonetics literatures (e.g., Hardcastle & Hewlett, 1999), with classic studies demonstrating that anticipatory coarticulation contributes to the identification of speech sounds (e.g., Kuehn & Moll, 1972; Martin & Bunnell, 1981). Several lines of work have investigated listeners’ sensitivity to coarticulatory information in spoken-word recognition.

Within-word coarticulation

Most studies examining within-word coarticulation have focused on what Whalen (1984) termed subcategorical mismatches. Building on a set of studies by Marslen-Wilson and Warren (1994; also see McQueen, Norris, & Cutler, 1999), Dahan, Magnuson, Tanenhaus, and Hogan (2001) cross-spliced the initial consonant and vowel of pairs of words such as *neck* and *net* to create either a combination of two tokens of the same word (e.g., *net* cross-spliced with another token of *net: [ne]t + ne[t]*) or a combination of different word tokens (*neck* cross-spliced with *net: [ne]ck + ne[t]*). For “different” word tokens, the formant transitions in the vowel preceding the word-final consonant in the token eventually

[1]We use the term “eye-movement control” following Altmann (2011) to include not only the programming and execution of saccades but also decisions to allow or cancel already programmed saccades. We note that regardless of whether the linguistic information results in programming of a new saccade or allowing or cancelling already programmed saccades, the effects will be reflected in the pattern of eye movements.

perceived as *net* were more consistent with the word *neck*. Beginning about 200 ms after the presentation of the initial consonant and vowel of the token *net*, with *[ne]* cross-spliced from *neck*, there were more fixations to a picture of a neck than to the eventual referent, a picture of a net (see Dahan & Tanenhaus, 2004, for related results). Given the 200-ms assumption, these results suggest that listeners used coarticulatory information in the vowel of the cross-spliced target word *net* that contained the vowel of the word *neck* to temporarily favor an interpretation of the speech signal as corresponding to the word *neck*, prior to hearing the final consonant of the word *net*.

Beddor, McGowan, Boland, Coetzee, and Brasher (2013) examined whether listeners use nasalization in a vowel to anticipate an upcoming nasal consonant. In American English, vowels are typically nasalized when followed by a nasal consonant, due to anticipatory lowering of the velum—especially when the vowel and nasal consonant occur in the same syllable. Participants saw a display with two pictures and heard a cross-spliced target word (e.g. *bent*) with nasalization starting early or late in the vowel. When the name of the competitor picture did not have a nasal consonant following the vowel (e.g., *bet*), participants' fixations converged on the target picture faster than when the name of the competitor also had a nasal consonant following the vowel (e.g. *bend*). Importantly, given the 200-ms assumption, this effect was apparent prior to processing of the segment following the vowel. Moreover, fixations converged on the target more quickly when nasalization occurred early in the vowel than when it occurred late in the vowel. Taken together, these results suggest that the dynamics of lexical activation closely follow the acoustic correlates of within-word (co)articulatory gestures in speech.

Pre word-onset coarticulation

Some evidence that pre word-onset coarticulation affects lexical activation comes from a study by Tobin, Cho, Jennett, Fowler, and Magnuson (in preparation). They used the printed-words version of the visual-world paradigm (McQueen & Viebahn, 2007; Salverda & Tanenhaus, 2010) with instructions such as "Pick up a pail"/"Pick up a pole", in which the realization of the indefinite article was influenced by the vowel in the following target word. Fixations to the target were delayed, and more fixations to the competitor were observed, when the indefinite article contained anticipatory coarticulation that did not match the upcoming vowel (e.g., [Pick up a] pail + Pick up a [pole] compared to a baseline cross-spliced sentence [Pick up a] pole + Pick up a [pole]). However, if we assume a 200-ms lag between speech and eye-movement control, the effects did not emerge until information from the vowel in the target noun was available.

Most other studies that have examined pre word-onset coarticulation have focused on English coronal place assimilation. For example, when a word that ends in an alveolar nasal (e.g., *lean*) is followed by a word that begins with a bilabial stop (e.g., *bacon*), the place of articulation of the nasal [n] partially assimilates to the place of articulation of the bilabial [b], resulting in an acoustic/phonetic pattern that is partially shifted towards an [m]. The literature on the processing of regressive place assimilation demonstrates that listeners make use of both the degree of assimilation in the speech signal and their knowledge about the environments in which assimilation occurs (Coenen, Zwitserlood, & Bölte, 2001; Gaskell, 2003; Gaskell & Marslen-Wilson, 1996, 2001; Gaskell & Snoeren, 2008; Gow, 2002). Most of these studies have provided limited information about the time course of effects of anticipatory coarticulation on word recognition because they have relied on tasks such as lexical decision, cross-modal priming, and phoneme monitoring. These methods require listeners to make a response based on a metalinguistic decision which takes about 500–800 ms. Most importantly, there is no direct way to link the timing of these responses to the processing of particular information in the speech signal.

The only evidence to date suggesting an immediate impact of pre word-onset anticipatory coarticulation on lexical access (and thus a stronger role for this type of information in spoken-word recognition) comes from a visual-world study by Gow and McMurray (2007) that focused on the processing of partially assimilated word-final coronal consonants. According to Gow's (2002, 2003) feature parsing model, partially assimilated segments combine elements of two places of articulation: that of the segment's underlying (i.e., coronal) form and that of the subsequent segment. The latter information can be used to anticipate the subsequent segment. Gow and McMurray used stimuli such as *green boat*, where *green* was either cross-spliced from an environment that would not result in place assimilation (e.g., *green dog*) or an environment where it typically undergoes place assimilation (e.g., *green bank*). In the assimilation environment, they found more fixations to the referent (*green boat*) within 120–160 ms after the onset of the word *boat*. Given the 200-ms assumption, these results would suggest that listeners used anticipatory coarticulation to anticipate the likely referent of the upcoming word.

Because place assimilation is arguably a special case of anticipatory coarticulation, it is not yet known whether or not anticipatory coarticulation prior to word onset will affect lexical access in a more general phonological environment. Therefore, we examined whether naturally occurring coarticulation allows listeners to anticipate an upcoming word when a determiner precedes a noun in a subject noun phrase (e.g., *The ladder is the target*). In order to answer this question it was also necessary to determine the minimal lag with which information in the speech signal can affect eye-movements.

How quickly does speech influence eye-movement control?

As noted earlier, most visual-world studies have assumed that the minimal lag is 200 ms. However, two recent studies suggest that it is important to reassess that assumption. Recall that the 200 ms assumption was motivated by results about the timing of visually-driven saccades. More recently, under conditions optimized for assessing the minimal lag for visually-driven saccades, Trottier and Pratt (2005) find evidence for effects earlier than 150 ms when participants are required to obtain information at the target location of the saccade. When this task is combined with a so-called sensory gap effect, where a central fixation cross disappears shortly before the target appears, saccade latencies as short as 70–80 ms were observed. Second, and most directly relevant to visual-world studies, Altmann (2011) has presented new analyses of two previously published studies that suggest that effects of speech on eye-movement control can be observed within 100 ms.

In both of the studies that Altmann reanalyzed (Altmann, 2004; Kamide, Altmann, & Haywood, 2003), participants saw a visual display depicting several objects and two or more potential agents. After a short delay, participants heard a prerecorded sentence. (In Altmann, 2004, the display had disappeared by that time, and eye movements to the locations previously occupied by the objects were analyzed.) In all of the materials, the initial noun phrase of the sentence referred to one of two potential agents in the visual scene. Each participant heard only one sentence associated with a particular visual scene, for instance, "*The man … will ride … the motorbike*" or "*The girl … will ride … the carousel*" in the context of a scene including a man and a girl. Therefore, each of the two potential agents served the role of target for one participant and the role of distractor for another participant.

The combined data from both conditions, within each experiment, was used to arrive at a principled estimate of the minimal lag at which information in the speech signal influences eye-movement control. Eye movements to the referent of the subject of the initial noun phrase are due to the linguistic signal and noise (that is, nonlinguistic factors, such as visual factors, and eye movement strategies employed by the participant or induced by the experimental task), whereas eye movements to the other potential agent are due to noise

only. Thus, the point in time where eye-movement control is sensitive to the processing of concurrently presented speech can be estimated by determining the earliest lag at which participants are more likely to initiate an eye movement to the target than to the other potential agent (hereafter, the associated distractor).

Altmann (2011) analyzed the latency of the first saccade launched after the onset of the target word. Latencies were sorted in time bins of 40 ms relative to the onset of the target word (which was determined by inspecting the speech signal). The first time bin in which the number of trials with a saccade to the target significantly exceeded the number of trials with a saccade to the associated distractor was 80–120 ms after the onset of the target word. This is considerably earlier than the 200-ms lag typically assumed (and frequently observed, e.g., by comparing looks to target and competitor pictures to looks to distractor pictures) in other visual-world studies. In none of these studies, however, were the data analyzed using Altmann's procedure.

To account for his results, Altmann (2011) proposed that the earliest effects of spoken language on eye movements in the visual-world paradigm might occur when pre-programmed eye movements are canceled. For instance, a planned eye movement to the man might be canceled upon hearing the initial sound of the word *girl*, because this information does not match the sound form associated with the picture of the man. We remain agnostic about whether or not the earliest effects of speech on eye-movement control might arise from inhibition of planned eye movements. Regardless of the mechanism responsible for the earliest effects, they might have been mediated by anticipatory coarticulation in the vowel that precedes the onset of the target word. If that were the case, the results could be consistent with a 200-ms delay.

In the data sets that Altmann analyzed, the spoken sentences all began with the definite article, *the*, immediately followed by a noun, spoken in a single phonological phrase, followed by a brief prosodic break. The target word was thus preceded by the unstressed central vowel /ə/ (schwa), which is strongly subject to coarticulation because it is produced with a neutral (that is, nearly unconstricted) vocal tract. Moreover, the vowel was preceded by the dental fricative /ð/, a speech sound which places few constraints on the position of the tongue body and would therefore be expected to only weakly affect the realization of the vowel. In addition, because the target word following the determiner was stressed, its initial sounds are expected to exert relatively strong coarticulatory effects (see Fowler, 1981; Magen, 1997, though note that these studies examined transconsonantal vowel-to-vowel coarticulation). Therefore, one would expect clear coarticulatory influences of the target word on the realization of the preceding determiner in Altmann's stimuli. Because the effects of speech on eye-movement control were assessed relative to the marked onset of the target word, anticipatory coarticulation might account for the earliness of Altmann's effects.

Altmann considered and rejected an explanation based on coarticulation for two reasons. First, he argued that if coarticulation were responsible for his early effects, then similar results should have occurred in other visual-world studies (e.g., Allopenna, Magnuson, & Tanenhaus, 1998). Second, he considered the timing of effects observed by Dahan et al. (2001), who as discussed earlier looked at effects of within-word anticipatory coarticulation, to suggest that coarticulation does not influence eye-movement control earlier than 200 ms. Note however, that the timing and magnitude of the effect of anticipatory coarticulation on eye-movement control would likely differ between studies, depending on the timing, strength and extent of the coarticulation in the relevant speech materials. In the previous paragraph, we discussed several reasons why one might expect that the anticipatory coarticulation preceding the target word would have been stronger and would have occurred earlier in Altmann's materials than in the studies he considered.

In the current studies, participants clicked on the referent of spoken stimuli, such as *The ladder is the target*, in the context of a four-picture visual display. We evaluated the hypothesis that listeners can use acoustic information in a determiner (*the*) preceding a spoken word (*ladder*) to develop lexical hypotheses about an upcoming word prior to the central articulation of its onset (here, the articulation of the sound /l/). We do so by first establishing the shortest latency at which eye movements are affected by processing of the onset of the noun when there is no anticipatory coarticulation pertaining to the target word in the preceding determiner (Experiment 1, e.g. *The … ladder is the target*). We then use this result as a baseline to determine whether the presence of anticipatory coarticulation preceding the noun results in shorter latencies (Experiment 2, e.g. *The ladder … is the target*).

Experiment 1

Method

Participants—Sixty-one students at the University of Rochester took part in the experiment. Participants were native speakers of American English, and reported normal hearing and normal or corrected-to-normal vision. We discarded the data from one participant because of a bad track. The remaining 60 participants ranged in age between 18 and 34 years, with a mean of 20.0 years. Forty-one of the participants were female.

Materials—Eighteen sets of four pictures were used in experimental trials. All pictures were selected from a widely used set of black-and-white line drawings developed and normed by Snodgrass and Vanderwart (1980). Each set consisted of two pictures that served the role of potential targets, and two distractors (see Fig. 1). We only used pictures with high name agreement scores (reported in Snodgrass & Vanderwart, 1980, as the percentage of participants who provided the name when shown the picture in a norming study), which we assumed would function as clear referents of the associated word. The average name agreement score for potential target pictures was 94.1 (out of 100); the average name agreement score for distractors was 88.6. A list of the words associated with the pictures in each of the stimulus sets is presented in Appendix A.

Each display included two potential targets. On a particular trial for a particular participant, one of these two pictures was mentioned in the spoken sentence. We refer to that picture as the target, and the other picture as the associated distractor. The associated distractor has a different status than the other two distractors included in the display, insofar as some of our analyses contrast looks to the target with looks to the associated distractor. Importantly, across participants, both potential targets equally often assume the role of target and associated distractor. Therefore, effects of purely visual factors on looks to a particular picture (for instance, some pictures may attract more looks than others irrespective of the spoken instruction) are effectively controlled for when comparing looks to target pictures to looks to associated distractors.

The onset of the name associated with each potential target picture was phonologically maximally distinct (in terms of articulatory features) from the onset of the names associated with the other three pictures in the set. Within a set, the initial sound of the names associated with the two potential target pictures (e.g. *ladder* and *fish* in Fig. 1) differed from each other in voicing, manner of articulation, and place of articulation; one of the names started with an approximant (/r/, /l/, or /w/) and the other with a voiceless fricative (/f/, /s/, or /ʃ/). The initial sound of the names associated with the two distractor pictures had a different manner and place of articulation from the initial sound of the name associated with each potential target. Importantly, the central, unstressed vowel schwa in the determiner preceding the target word is likely to undergo anticipatory vowel-to-vowel coarticulation (Fowler, 2005;

though note that in the current experiment, this effect is expected to be weaker than anticipatory coarticulation due to the onset of the target word; cf. Cole, Linebaugh, Munson, & McMurray, 2010). The onset of each of the four words associated with a set of pictures was therefore followed by a different vowel. The onset of the spoken target word, and in particular anticipatory coarticulation associated with the target word, should thus provide minimal support for the referent being any of the pictures other than the target picture.

An additional 11 sets of four pictures were used in practice and filler trials. The names associated with the (potential) targets in these sets started with the sound /h/, /v/, /m/, /p/, /t/, /k/, /b/, /g/ or /tʃ/. These trials were included to discourage participants from expecting that target words were likely to start with an approximant or a voiceless fricative (as was the case for all experimental trials). Six filler sets included two potential targets (that is, the target varied across participants), and five practice sets included a fixed target. Each set of pictures was constructed in accordance with the same set of phonological criteria used for experimental trials, with the exception that on two practice trials, one distractor was associated with a name whose initial sound had the same manner of articulation as the initial sound of the target.

Spoken sentences were recorded digitally in a sound-treated booth by a female speaker of American English, using 44.1 kHz sampling rate and 16 bit amplitude resolution. The order of sentences was randomized. Each target word was recorded twice, once in the sentence frame used for Experiment 1 (e.g., “The … ladder is the target”) and once in the sentence frame used for Experiment 2 (e.g., “The ladder … is the target”). In the stimuli used in Experiment 1, the determiner and the target word were separated by a brief prosodic break. In the stimuli used in Experiment 2, the determiner and the target word constituted a continuous phonological phrase, which was followed by a brief prosodic break (following the phrasing of the initial noun phrase in the stimuli used in Kamide et al., 2003). The speaker first recorded all of the sentences for Experiment 2, followed by all of the sentences for Experiment 1, using the same randomized order of sentences. During the latter part of the recording, she also mixed in 11 tokens of the word “the” in isolation, pronounced with a similar intonation as at the start of the sentences.

One token of the word “the” recorded in isolation was selected to be cross-spliced at the onset of all sentences recorded for Experiment 1. The average duration of the word “the” at the onset of these sentences was 332 ms, and that of the following prosodic break 225 ms. Therefore, we selected a token of “the” recorded in isolation of 385 ms (which most closely resembles the average duration of the determiners in the sentences), followed by 225 ms of the prosodic break following that token—which consisted of minimal acoustic energy in the form of the background noise recorded during the prosodic break. This material was then used to replace the word “the” and the prosodic break originally recorded with each sentence. As a result, each of the sentences used in Experiment 1 began with the same token of “the”, which was not coarticulated by any following speech material, followed by a prosodic break of 225 ms.

The first author used the speech editor Praat (Boersma &Weenink, 2012) to label the onset and offset of the determiner and the target word in each of the 36 experimental sentences recorded for Experiment 1. This was done on the basis of auditory information in the speech signal in combination with information derived from a visual representation of the speech signal in the form of a waveform and a spectrogram. The point in time where the speech signal following the prosodic break first showed evidence of vocal activity was marked as the onset of the target word. After the experiment had been completed, two trained phoneticians independently determined the onset of vocal activity associated with the target word in each sentence. Between labelers, there was relatively little variation in word-onset

estimates. The average difference between any two labelers was 8 ms for targets starting with a voiceless fricative (ranging from 0 to 32 ms; *SD* = 9 ms), and 9 ms for targets starting with an approximant (ranging from 0 to 82 ms; *SD* = 15 ms).

Word onset estimates of the first author were used to create the cross-spliced sentences. That is, in those cross-spliced sentences the one selected token of the word "the" and the following prosodic break were followed by the target word starting from the point in time where the first author had determined that the target word started. For 24 of the 36 sentences, the first author's estimate of word onset corresponded to the median onset estimate among the three labelers. For the remaining 12 sentences, the median onset estimate among the three labelers was earlier than the first author's estimate, by 11 ms on average. However, due to the fact that the cross-spliced sentences had been created on the basis of just the first author's judgment, and used as such in the experiment, it was not possible to take measurements from all three labelers into account in analyzing the data. According to the judgments of the other two labelers, for 12 sentences, a very small part of the target word was missing at its onset. However, the onset of a word preceded by silence is typically relatively quiet (with the exception of words starting with a plosive or a glottal stop), and the cross-spliced words sounded completely natural in their sentence contexts.

Design—We minimized the potential impact of visual factors on visual attention and eye movements by counterbalancing, across displays, the location of the two potential targets, and which of the two potential targets was the target and which was the associated distractor. The position of the pictures in each set was randomized with the constraint that each of the six possible constellations of two pictures that were potential targets (e.g., one in top-left position and one in bottom-right position) occurred four times within the combined set of experimental and filler trials (24 in total). Across four lists, we manipulated which of the two potential target pictures was the designated target, and in which of two locations it appeared. For instance, if the potential targets *ladder* and *fish* were associated with the top-left and bottom-right location in the associated visual display, the spoken word was "ladder" on two of the four lists, with the ladder occurring in the top left corner (and the fish in the bottom-right corner) on one of those lists, and with reversed picture locations on the other list; the spoken word was "fish" on the other two lists, with the fish occurring in the top left corner (and the ladder in the bottom-right corner) on one of those lists, and with reversed picture locations on the other list. Within each list, the target word began with a voiceless fricative on half of the experimental trials, and with an approximant on the other half of the experimental trials.

Fifteen random trial orders were created and combined with each list, resulting in a total of 60 randomized lists. The same set of five practice trials was appended at the start of each randomized list, and each participant was assigned to one of the randomized lists.

Procedure—Eye movements were monitored using an EyeLink II head-mounted eye-tracking system, sampling at 250 Hz. The eye-tracker was calibrated prior to the experiment. Each trial began with the presentation of a 3 × 3 grid in the center of the screen, which included four pictures, one in each corner cell of the grid (see Fig. 1). After a delay of one sec, a spoken instruction was presented through a pair of Sennheiser HD 570 headphones. After the participant clicked on a picture in the visual display, the screen went blank for 2.5 s before the start of the next trial. Every five trials, a small dot appeared in the center of the screen. The participant was required to fixate and click on the dot in order to perform a drift correction for the eye-tracking system.

The eye-tracker data were parsed into a sequence of events using Eyelink software, which uses thresholds for angular eye movement, velocity, and acceleration to classify streams of

gaze coordinates into a sequence of saccades, fixations, and blinks. We used custom-made software designed by Edward Longhurst to automatically code fixations to the objects in the visual display. A saccade was coded as directed to the location subsequently fixated. Fixations within a cell of the grid in which a picture was presented, or within an area just outside the grid cell extending to one sixth of its width, were coded as fixations to that picture. Fixations were coded as directed to the target, the associated distractor, the other two distractors, or any other location on the screen.

Results

Fixation proportions—For each 4-ms time bin, we computed the proportion of fixations to each of the pictures in the visual display. Fig. 2 presents these fixation proportions, averaged across participants and aggregated in 20-ms time bins.

At the onset of the target word, the proportion of fixations to the target was slightly higher than that to the associated distractor, and the proportion of fixations to the associated distractor was slightly higher than that to the averaged other distractors. Importantly, the difference in the proportion of fixations to the target and associated distractor was not significant (in a logistic mixed effects models with random intercepts for participants and items, $p = .44$).[2] This suggests that the differences in fixation proportions to target and associated distractor at word onset were due to chance, a conclusion that is further supported by the fact that no such differences in fixation proportions were found in Experiment 2, which used the same set of target words and visual displays. Crucially, there were no apparent changes in fixation proportions to either type of picture between 0 and 200 ms after the onset of the target word. Shortly after 200 ms, fixations to the target began to increase rapidly while fixations to the associated distractor and the other distractors began to decrease. Fixations to the three distractors did not increase above their baseline levels.

Saccade latencies—Following the procedure introduced by Altmann (2011), we analyzed saccade latencies to the target and associated distractor, focusing on just those trials on which the first saccade following the onset of the spoken word was launched to the target or associated distractor. This necessarily required looking at just a subset of the trials.

Trials were excluded from subsequent analysis for the following reasons: no saccade was generated to any of the pictures after the onset of the target word (2.3% of the data); the first saccade following the onset of the target word was directed to a picture that the participant was already fixating (23.9% of the remaining data); the first saccade following the onset of the target word was not directed to either the target or the associated distractor (50.6% of the remaining data). Saccade latencies were computed for the remaining trials (36.8% of the data) and sorted in 40-ms bins, relative to the onset of the target word.

Fig. 3 presents a histogram of saccade latencies to the target and the associated distractor. Prior to 200 ms after the onset of the target word, there were no clear numerical trends in any of the 40-ms time bins. Beginning at the 200–240 ms time bin, the number of trials with a saccade to the target always exceeded the number of trials with a saccade to the associated distractor.

Following Altmann (2011), we performed Chi-Square tests comparing the number of saccades to the target and associated distractor in 40 ms time bins.[3] In order to assess as

[2]Specifically, we examined the odds of fixating the target relative to the associated distractor at the onset of the target word. Only trials on which one of the relevant events occurred were included in the analysis. In a logistic mixed effects model, with random intercepts for participants and items, the log odds of fixating the target over the associated distractor (i.e., the intercept of the model) was not different from zero ($\beta = .080$, 95% CI [−0.12, 0.28]; note that $\log(1/1) = 0$).

precisely as possible at what point in time participants were more likely to generate a saccade to the target than to the associated distractor, we used a moving-window analysis. This analysis uses the same bin size as examined in Altmann's (2011) analysis. However, the window moves by increments of 10 ms, which allows for a more fine-grained analysis of the point in time at which processing of the speech signal influences fixation patterns. The moving window starts at word onset and moves to 400 ms after word onset, in 10 ms increments. At each of those points in time, we computed the number of trials with a saccade to the target and the number of trials with a saccade to the associated distractor in a window extending 40 ms in time. Thus, the first window included saccades generated between 0 and 40 ms after word onset, the second window included saccades generated between 10 and 50 ms after word onset, etc.

The results of the moving-window analysis are presented in Fig. 4. Up to 200 ms after the onset of the target word, participants were not significantly more likely to generate a saccade to the target than to the associated distractor in any of the 40 ms time bins tested. The earliest windows during which participants were found to be significantly more likely to generate a saccade to the target than to the associated distractor corresponded to 200–240 ms and 210–250 ms after word onset. This corresponds to when looks to the target and the distractors begin to diverge in the visualization provided by the proportion of fixations curves (see Fig. 2). The effect was also significant for later time bins, starting around 300 ms and 400 ms after word onset. Importantly, our moving-window analysis did not provide any evidence for effects of speech on eye-movement control earlier than 200 ms after the onset of the target word.

Discussion—The goal of Experiment 1 was to establish a baseline for the time it takes for information in the speech signal to influence eye movements to potential referents in a visual-world experiment with a four-picture display. Since the target word was preceded by a determiner followed by a brief prosodic break, which had been cross-spliced from a neutral context, the speech signal did not contain any anticipatory coarticulatory information pertaining to the onset of the target word. Moreover, because the initial sound of the target word strongly mismatched the initial sound of the names associated with the other pictures in the visual display, fixations to the target should be minimally affected by the consideration of other pictures in the display on the basis of information in the speech signal. Given the criteria used for the selection of pictures, the pairing of two potential target pictures, the differences in sound form between the onsets of the names associated with the pictures within each visual display, the controlled acoustic environment, and our experimental design, we created circumstances in which the influence on eye-movement control of factors other than the match between the spoken target word and the target picture was expected to be minimal. The pattern of fixation proportions over time and the analysis of latencies of the first saccades following the onset of the target word converge on an estimate of 200–240 ms. One possible concern is that Experiment 1 might not have had sufficient power to find significant effects of the speech signal on eye-movement control prior to 200 ms. Power analyses which demonstrate that this is unlikely are presented in Appendix B.

[3]The smaller the analysis window, the more precisely one can estimate the onset of the effect of speech on eye movements. However, given the relatively low number of trials contributing to the analysis, the data were too sparse to use shorter analysis windows. Note that even with 40-ms time bins, it is not possible to fit a logistic mixed effects model with random effects for participants and items to the data, because there are too many empty cells (that is, participants or items for which there are no trials on which a saccade was initiated toward the target or associated distractor in a particular time bin).

Experiment 2

We are now in a position to examine the effects of anticipatory coarticulation on the generation of lexical hypotheses. In Experiment 2, a determiner and target word were presented in a single phonological phrase, which was followed by a brief prosodic break (e.g., *The ladder … is the target*). As discussed earlier, the vowel in the utterance-initial determiner was expected to contain information about the onset of the following word. The question, then, is whether listeners can use such information to facilitate recognition of the target word, and, if so, how quickly anticipatory coarticulation influences the recognition of an upcoming spoken word.

Method

Participants—Sixty-five students at the University of Rochester took part in the experiment. All were native speakers of American English, and reported normal hearing and normal or corrected-to-normal vision. Data from three participants were discarded because they made no eye movements following the onset of the target word on more than 25% of experimental trials (28%, 39% and 67%, respectively). In addition, data from one participant were discarded because we identified a number of irregular timing messages in their eye-tracker output, and data from another participant were discarded because of a bad track.

The remaining 60 participants ranged in age between 18 and 24 years, with a mean of 19.7 years. Forty-five of the participants were female.

Materials—Due to the continuous nature of the speech signal and the prevalence of coarticulation in the sentences recorded for Experiment 2, the onset of the target word can only be estimated. The first author used the following criteria to mark the onset of the target word. When the target word began with a voiceless fricative, word onset was defined as the point in time where the speech signal was first associated with high-frequency frication noise, as assessed from a waveform and spectrogram representation of the speech signal. When the target word started with an approximant, word onset was defined as the point in time where a sustained region of low spectral energy across formants started (in particular, a reduction in spectral energy for F3 and F4).

In addition to the first author, two trained phoneticians were asked to label the onset of the target word in each sentence. They were not provided with instructions about how to segment the speech signal. For each item, the median measurement of word onset among the three labelers was selected as the best estimate of word onset. Across labelers, there was relatively little variation in estimates. The average difference between any two labelers was 5 ms for target words beginning with a voiceless fricative (ranging from 0 to 13 ms; *SD* = 4 ms), and 8 ms for target words beginning with an approximant (ranging from 0 to 25 ms; *SD* = 7 ms).

Analysis of stimuli—In order to determine if, and if so, when, there was sufficient information in the determiner for a listener to, in principle, predict the next segment (at a level above chance), we first performed an acoustic analysis of the formant structure of the schwa preceding the target word. We then trained a Gaussian classifier to predict the onset of the target word on the basis of the formant center frequencies associated with the schwa, measured at different points in time in the determiner.

Acoustic measurements—We measured the frequencies of the first three formants of the vowel schwa, which immediately preceded the target word. The mean duration of this vowel was 53 ms (42 ms when it preceded a fricative, and 63 ms when it preceded an

approximant). The formant structure of speech sounds reflects aspects of their articulation, and these relationships have been well established in the phonetics literature. The first formant (F1) reflects the distance between the tongue body and the palate, that is, the height of the tongue during the articulation of a sound: higher F1 is associated with a lower tongue position. The second formant (F2) reflects the front–back positioning of the tongue body, as well as the degree of lip rounding: fronting the tongue body raises F2 and retracting the tongue body lowers F2, while lip rounding also lowers F2. The third formant (F3) has a more complex relationship with articulation. Importantly, for the purposes of the present study, F3 is affected by the gestures that are characteristic for English approximants, and differentially so for the approximants that we focus on in the current study (/l/, /r/, and /w/). Because schwa is an unstressed central vowel, produced with a neutral vocal tract, we expected that its quality (and, at an acoustic level, its formant structure) would be influenced by the articulation of the initial sound of the immediately following target word. A priori, it is not clear how early in the vowel coarticulatory information pertaining to the upcoming noun might be present. We therefore measured the center frequencies of the first three formants at the onset, midpoint, and offset of the schwa, and examined how well our classifier was able to predict the identity of the initial sound of the upcoming target word on the basis of each of these sets of three (F1, F2, and F3) measurements.

The first author used the speech editor Praat to measure, for each speech stimulus, the center frequencies of the first three formants at the first pitch period of the vowel (that is, the earliest point in time where the vowel's formant structure could be measured, as evidenced by the presence of formant bands in the spectrogram), half-way into the vowel, and at the offset of the vowel (the point in time which had been marked as the onset of the target word on the basis of judgments from three labelers). Praat computes formant frequencies by using linear predictive coding (LPC) over a Gaussian-like analysis window that is centered on the point of measurement. The number of LPC coefficients was set to 10; twice the maximum number of formants (which was set to 5). Maximum formant frequency was set to 5500 Hz, which is used conventionally for female speakers. We constrained the duration of this window as much as possible; it was 20 ms long, with 96% of the measurement mass falling in the center 10 ms. Our female speaker's fundamental frequency at the onset of schwa was about 220 Hz, and one pitch period of the vowel thus corresponded to between 4 and 5 ms. Therefore, formant measurements at the onset of the schwa were influenced by information in the first few pitch periods of the vowel, as well as the final few pitch periods of the preceding voiced dental fricative.

Fig. 5 presents formant frequencies for F1, F2 and F3, at the onset, mid point and offset of the vowel, averaged across items sharing the same following sound. There are clear differences in formant trajectories across schwa as a function of the following phoneme. F1 drops at schwa offset, reflecting the upward movement of the jaw involved in narrowing of the vocal tract for approximants, or even further narrowing of the vocal tract to produce partial airstream obstruction for fricatives. F2 rises throughout schwa for the sibilants /s/ and /ʃ/, reflecting forward movement of the tongue to the alveolar ridge and the forward part of the hard palate, respectively. F2 drops markedly for the approximants, reflecting backward movement of the body of the tongue. F3 is relatively flat throughout schwa for the fricatives, but shows a distinct pattern for each of the approximants. There is a sharp drop in F3 preceding /r/, which is typical for retroflex movement; a smaller drop in F3 preceding /w/, which is typical for lip rounding; and a rise in F3 preceding /l/, which is characteristic for English /l/. We note that these patterns are well established in the acoustic phonetics literature.

Although differences in formant frequencies as a function of the upcoming consonant are most apparent in the formant trajectories across schwa, even the formant frequencies at the

very onset of the vowel show some systematic signs of anticipatory coarticulation. Generally speaking, formant values at the onset of schwa anticipate the direction of formant movement observed throughout schwa: At schwa onset, F2 tends to be high when it rises throughout schwa (e.g. for /s/ and /ʃ/), and low when it drops throughout schwa (e.g. for the approximants). F3 shows a similar trend for the approximants, compared to the steady level of F3 at schwa onset across the fricatives— which show little movement of F3 throughout the vowel compared to the approximants. Thus, there is some indication that effects of anticipatory coarticulation of the following consonant are apparent at the onset of the vowel in the determiner.

Classifier results—In order to determine if—and when—there is sufficient information in the formant center frequencies to predict the initial sound of the upcoming word, we used a classifier analysis to predict the initial sound of the target word from the center frequencies of the first three formants at the onset, midpoint, or offset of the vowel in the determiner preceding the target word. In our stimuli, the onset of the target word corresponded to one of six phonemes: the fricatives /f/, /s/, and /ʃ/, and the approximants /r/, /l/, and /w/ (see Appendix A). For the classifier analysis, these initial sounds serve as category labels for the formant values from the different schwa tokens. We used a Gaussian classifier to assign category labels based on the distance from each category, where the distance function depends on the covariance between the three formants from the set of schwa tokens used to train the classifier. In our case, training the classifier meant determining, for each onset category, the mean of and covariance between the three formants of the schwa tokens belonging to that category. The classifier was then used to predict category labels (i.e., the initial sound of the following word) of a test token by calculating how similar the test token's formant values were to each category, where similarity was determined by the likelihood of the test token's formant values under a three-dimensional Gaussian distribution with the mean and covariance for each category as determined during training. The output is thus, for each test token, the probability of that test token belonging to each onset category. We discarded the two stimuli starting with /ʃ/ because a non-singular three-dimensional variance–covariance matrix cannot be computed on the basis of only two tokens. The classifier was trained using a leave one-out cross validation technique: for each token, the classifier was trained using all the *other* tokens.

At the onset of the vowel in the determiner, that is, using formant frequency information measured at the vowel's first pitch period, classifier performance is 50%, which is much better than chance.[4] Performance improves to 88% at the midpoint of the vowel, and drops slightly, to 85%, at the end of the vowel.[5]

A confusion matrix representing output from the three classifiers is presented in Fig. 6. Each cell represents the probability of assigning a token to each of the five possible categories (/r/, /l/, /w/, /f/, and /s/), with brighter hues representing higher likelihoods. Even the classifier trained using the formants at schwa onset is most likely to correctly assign tokens from a particular onset category to their category—with the exception of /r/-initial tokens, which are more likely to be classified as /w/ than as /r/. The midpoint and offset classifiers assign tokens to the correct category with very high likelihood, with close to uniform performance across the five onset categories. Importantly, each classifier shows good performance for

[4]Classifier chance accuracy was determined by a permutation test, randomly shuffling the onset labels of the tokens and following the same training and test procedures as described above. This procedure was repeated 100 times to obtain an estimate of the classifier's performance under the null hypothesis that the onset of the following word bears no relationship with the tokens' formant values. By this test, chance accuracy of the classifier was 19% (95% CI [6%, 35%]).

[5]We also ran the classifier after converting formant frequencies to the Bark scale, which reflects a better approximation of the human perception of frequency—which is non-linear. Classifier performance was very similar: 47% at vowel onset, and 85% at the midpoint and at the end of the vowel. The associated confusion matrix was virtually identical to the one presented in Fig. 6.

fricatives as well as approximants. This suggests that average classifier performance is not exclusively due to good predictions for the approximant tokens, which were associated with greater movement in F2 and F3 throughout the schwa than the fricative tokens.

The classifier results demonstrate that within the set of experimental sentences in Experiment 2, even the first few pitch periods of the determiner's vowel contain information that a listener could, in principle, use to predict the initial sound of the following target word. (Note that coarticulatory information pertaining to the target word may also be present in the determiner's initial sound, the voiced dental fricative /ð/. However, such coarticulatory information is expected to be relatively weak and is hard to measure or quantify acoustically.) Using the 200 ms baseline established in Experiment 1, we now examine if, and if so, when, listeners use anticipatory coarticulatory information.

Design and procedure—The design, procedure, and randomized lists of trials were identical to those used in Experiment 1.

Results

Fixation proportions—Fig. 7 presents fixation proportions to each type of picture, relative to the onset of the target word and averaged across participants. At word onset, participants were equally likely to fixate either of the pictures in the display. Between 100 and 200 ms after word onset, fixations to the target began to increase, while fixations to the associated distractor and the other two distractors began to decrease. Fixations to the distractors never increased above their baseline level. Thus, speech-driven fixations to pictures in the display occurred within 200 ms after the marked onset of the target word, suggesting that listeners made immediate use of anticipatory coarticulation.

In order to evaluate the time course of the effect of anticipatory coarticulation on the recognition of the target word, we directly compared the time course of fixations to the target between Experiment 1 and Experiment 2 (see Fig. 8). Fixations to the target increase earlier, relative to the marked onset of the target word, in Experiment 2, where the target word was preceded by anticipatory coarticulation. This demonstrates that listeners used anticipatory coarticulation very early in the recognition of the target word.

Once target fixations began to rise, compared to their baseline at target-word onset, they increased at a faster rate in Experiment 1 than in Experiment 2. Target fixations also reached a higher plateau in Experiment 1 compared to Experiment 2. This effect goes in the opposite direction of the effect of anticipatory coarticulation that is evident very early in the processing of the target word. We note that there were durational differences between the target words in the two experiments, due to differences in the prosodic phrasing of the instructions. We would expect target words in Experiment 2 to be longer because they were followed by a prosodic break and therefore associated with phrase-final lengthening (Oller, 1973). The mean duration of the target words was indeed longer in Experiment 2 (541 ms) than in Experiment 1 (419 ms). A paired two-tailed *t* test revealed that this difference was significant, $t(35) = 8.7$; $p < .001$. As a consequence, information following the onset of the target word accrued more quickly in Experiment 1 than in Experiment 2. When segmental information pertaining to the sound form of the target word becomes available earlier, evidence for the target word is expected to increase more rapidly. Thus, the difference in the rate of increase of target fixations between the two experiments likely reflects differences in the uptake of phonetic information as the target word unfolds, which are due to systematic differences in duration of the target word between experiments. Importantly, the difference in the rate of increase in target fixations occurred after the effect of anticipatory coarticulation, which is of primary interest in this study.

Saccade latencies—We only analyzed trials on which the first saccade following the onset of the target word was initiated towards the target or associated distractor. Trials were discarded from the analysis for the following reasons: no saccade was generated to any of the pictures after the onset of the target word (2.0% of the data); the first saccade following the onset of the target word was directed to a picture that the participant was already fixating (20.5% of the remaining data); the first saccade following the onset of the target word was not directed to the target or the associated distractor (49.9% of the remaining data). Saccade latencies were computed for the remaining trials (39.0% of the data) and sorted in 40-ms bins, relative to the onset of the target word.

Fig. 9 presents a histogram of saccade latencies to the target and the associated distractor. Prior to 120 ms after the onset of the target word, there were no clear numerical trends in any of the 40-ms time bins. Beginning at the 120–160 ms time bin, the number of trials with a saccade to the target always exceeded the number of trials with a saccade to the associated distractor, with the exception of the 240– 280 ms time bin.

The results of a moving window analysis are presented in Fig. 10. The earliest window during which participants were significantly more likely to generate a saccade to the target than to the associated distractor corresponded to 130–170 ms after word onset. (Note that a marginally significant effect was observed in the 110–150 ms and 120–160 ms time bins.) From there on, the effect persisted until the 200–240 ms time bin. The effect was also significant for the majority of time bins starting between 300 and 400 ms after word onset. The moving window analysis therefore provides clear evidence that when the determiner that precedes the target word is associated with anticipatory coarticulation pertaining to the target word, effects of speech on eye movements occurred within 200 ms of the (marked) onset of the target word. Importantly, this effect occurred about 70 ms earlier in Experiment 2 than in Experiment 1, where no anticipatory coarticulation pertaining to the target word was available prior to its onset. This suggests that listeners use anticipatory coarticulation to predict the initial sounds of an upcoming spoken word.

If participants used anticipatory coarticulation at the onset of the determiner's vowel to anticipate the target word, we would expect that if fixations are aligned to the onset of the vowel in the determiner rather than to the onset of the noun, then fixation proportions to the target should rise with the same delay as was observed in Experiment 1. Fig. 11 presents fixation proportions to each type of picture in Experiment 2, aligned to the onset of the vowel in the determiner (e.g. in the sequence *The ladder* …). Fixations to the target begin to rise at or very shortly before 200 ms after the onset of the schwa, while fixations to the associated distractor and the other distractors drop. Taking into account that it takes about 200 ms for speech to influence eye-movement control, as established in Experiment 1, the data suggest that participants made immediate use of anticipatory coarticulation in the vowel. Converging evidence comes from a comparison of target fixations across time between Experiment 1, aligned to the onset of the target word, and Experiment 2, aligned to the onset of the vowel of the determiner preceding the target word (see Fig. 12). Target fixations begin to rise about 200 ms after the alignment point in both experiments.

Taken together, the results from Experiment 2 demonstrate that listeners make rapid use of anticipatory coarticulation during spoken-word recognition. Comparison of the data from Experiments 1 and 2 also suggests that fixations to potential referents in a standard four-picture version of the visual-world paradigm first reflect the uptake of phonetic information with a delay of approximately 200 ms. Unlike Altmann (2011), we did not find significant effects as early as 80–120 ms after word onset. However, the exact timing of effects in the two studies cannot be compared directly. There are differences in stimuli, design, procedure, and experimental task between our study and Altmann's that might account for differences

in results. Most crucially, from our perspective, a comparison between the two studies would require realigning the fixation data in Altmann's studies to the onset of the earliest reliable anticipatory coarticulation in the determiner in his stimuli. We note that the determiners in Altmann's studies were longer than the determiners in our studies. The mean duration of the determiner was 176 ms in Kamide et al. (2003) and 170 ms in Altmann (2004), whereas it was 100 ms in the current study (ranging from 58 to 169 ms). If we make the assumption that the degree and relative temporal extent of anticipatory coarticulation in Altmann's stimuli and ours were similar, then information was available to the listener earlier, relative to the (marked) onset of the target word, in Altmann's stimuli. This would then explain why Altmann found effects in the 80–120 ms time bin following the onset of the spoken word, whereas our earliest effects were in the 130–170 ms bin.

General discussion

We used the visual-world paradigm to examine the role of pre word-onset anticipatory coarticulation in the recognition of spoken words. Using instructions such as "The … ladder is the target", we first established that when the determiner did not contain anticipatory coarticulation pertaining to the target word, the earliest effects of speech on eye-movements occurred between 200 and 240 ms after the onset of the target word (Experiment 1). We then used a Gaussian classifier to demonstrate that, for naturally produced utterances in which the determiner and noun were produced as one phonological phrase (e.g., "The ladder … is the target"), coarticulatory information was present in the first few pitch periods of the determiner's vowel, and continued throughout the vowel. With these utterances, the earliest signal-driven fixations to the target picture began between 130 and 170 ms after the onset of the target word—about 70 ms earlier than in Experiment 1. The shift in timing maps closely onto the average duration of the vowel of the determiner preceding the target word, which was 53 ms and ranged from 30 to 78 ms. Taken together, these results demonstrate that listeners made immediate use of anticipatory coarticulation in the determiner to predict the initial sound(s) of the upcoming word.

We first discuss the methodological implications of our results for visual-world experiments and then discuss the theoretical implications for models of spoken-word recognition, returning to the issues we raised briefly in the introduction.

Methodological implications for visual-world experiments

Researchers using the visual-world paradigm typically align fixations to theoretically-defined regions of interest in the speech signal, assuming that it takes about 200 ms for information in the speech signal to influence eye movements to potential referents. Our results provide clear support for this assumption. Under conditions created to detect the earliest effects of speech on eye-movement control in a four-picture display, we found no evidence for such effects within 200 ms after target-word onset. We believe that our estimate of the lag is likely to correspond to the minimum time it takes for speech to influence fixations in a task-based version of the paradigm with a four-picture display. In Experiment 1, the target word was clearly articulated, phrase-initial, and preceded by a short period of silence (thus, minimizing the requirements for segmentation). The visual context consisted of a small array of objects from a set of normed pictures, which appeared in fixed locations on the screen. Moreover, there was no phonemic overlap (and carefully controlled, minimal featural overlap) between the initial sounds of the target word and those of the names associated with each of the pictures within each display. Finally, because we used the same token of "the" followed by a prosodic break of a constant duration, the onset of the target word was predictable. All of these factors conspire to create a situation in which the mapping of the spoken word onto the target picture should be as straightforward as possible.

Having established that the earliest effects of speech on eye-movement control occur with a lag of about 200 ms allows us to reevaluate the time course of effects of coarticulation in a number of previously discussed visual-world studies. The results from Gow and McMurray (2007) are indeed evidence for immediate effects of pre word-onset anticipatory coarticulation due to place assimilation on the recognition of a following word. Moreover, the timing of the effects observed in Beddor et al. (2013), on nasalization, and Dahan et al. (2001), on sub-categorical mismatches, suggest that listeners rapidly use within-word anticipatory coarticulation, prior to the processing of subsequent speech material. Finally, the results support the prevalent 200 ms minimal lag assumption, which has played a central role in the interpretation of many results in higher-level language processing (e.g., the line of research initiated by Sedivy et al., 1999, which we discussed earlier).

Our experiments were not designed to duplicate the conditions from the experiments reanalyzed in Altmann (2011).We acknowledge that aspects of the design of those experiments—for example, the use of the "look and listen" paradigm, or characteristics of the visual displays, such as the fact that there were only two potential referents for the noun phrase in many trials—could have contributed to the earlier effects of speech on eye-movement control reported by Altmann. With those caveats aside, we think that the most likely explanation for Altmann's lower-bound estimate of a 100 ms lag is that reliable anticipatory coarticulation preceded the estimated onset of the noun by about 80–100 ms in his stimuli. If that was indeed the case, then Altmann's results are consistent with a 200 ms lag between information in the speech signal and the first effects of that information on eye-movement control.

Given the ubiquity of coarticulation in speech, and the close correspondence between what we observe in proportion of fixation plots and more detailed time-course analyses, one might wonder why reports of effects earlier than 200 ms have not been routinely observed in the visual-world literature. The most likely reason is that most visual-world experiments do not have the combination of factors necessary to observe these early effects of anticipatory coarticulation (e.g., the presence of relatively strong and temporally extensive anticipatory coarticulation preceding the target word). Nonetheless, when inferences depend crucially on the argument that fixations preceded some designated landmark in the speech stream it will be important to report more details about the speech signal and how word onsets were estimated than has been common practice in the visual-world literature.

Finally, our results confirm the usefulness of proportion-of-fixation plots, introduced in Allopenna et al. (1998), as a data visualization tool for visual-world studies. In both experiments there is a clear correspondence between when effects are first observed in plots presenting proportion of fixations over time and when effects become reliable in the moving window analyses of first saccades following the onset of the target word.

Implications for models of spoken-word recognition

In most models of spoken-word recognition, word recognition has either explicitly or implicitly been conceptualized as a pattern-recognition problem in which acoustic/phonetic information activates stored linguistic representations (e.g., features, phonemes, and words) that compete for activation (Marslen-Wilson, 1987; McClelland & Elman, 1986; Norris, 1994) or probability space (Norris & McQueen, 2008) based on their goodness of fit with the speech signal. In these models, word recognition is closely time-locked both to the onset of evidence for a word in the speech stream as it unfolds over time and the goodness of match of the input to its competitors (i.e., similar sounding words). These models could incorporate effects of anticipatory coarticulation by broadening the window in which possible evidence for lexical candidates might occur (see also Elman & McClelland, 1986). Thus, rapid effects of pre word-onset anticipatory coarticulation are not inconsistent with

these models, though the models may provide limited insight into when those effects will arise. Moreover, it is not a priori clear that the current input representations of those models, for instance the featural representations in TRACE (McClelland & Elman, 1986), could capture the temporal extent of coarticulation that we observed in the current study without compromising the performance of the model. Shortlist B (Norris & McQueen, 2008) acknowledges the role of coarticulation in spoken-word recognition by using probabilistic phonemic input representations that are based on listener identifications of diphones in a largescale gating study (Smits, Warner, McQueen, & Cutler, 2003). Although such information improves the performance of the model, this solution bypasses the modeling of one level of processing of the speech signal. Importantly, as we argue below, how listeners process coarticulatory information may give important insights into fundamental aspects of the workings of the spoken-word recognition system.

Exemplar models, such as Goldinger (1998), Johnson (1997), and Pierrehumbert (2002) could naturally account for, and indeed would predict, effects of fine-grained acoustic information, including within-word coarticulation. The easiest way of incorporating pre word-onset anticipatory coarticulation is to assume that (some) multi-word units, for example sequences such as determiners followed by nouns, are stored as exemplars. A determiner noun exemplar would clearly encode information about the upcoming noun that is present in the vowel in the determiner. However, if we assume that (more subtle) effects of anticipatory coarticulation influence the processing of a given word in a wide variety of lexical contexts, exemplar models would need to store an enormous number of different multi-word sequences to account for those effects. If exemplar models store only a limited number of multi-word sequences, or if the quality of those representations differs as a function of how frequently a listener has encountered a particular multi-word sequence, exemplar models would predict that the influence of anticipatory coarticulation on the recognition of a following word would depend on the frequency with which the listener has encountered the two-word sequence. The effect of anticipatory coarticulation would be predicted to be strongest for two-word sequences that occur with relatively high frequency. According to data-explanation approaches, predictability should be the controlling variable and not frequency, as would be predicted by exemplar models. The frequency of two-word sequences is likely to be highly correlated with the predictability of the second word given the first. Nonetheless, it should be possible to distinguish between the exemplar and data-explanation accounts by independently manipulating frequency and predictability in an experimental paradigm.

In the last few decades, data-explanation frameworks have emerged as a theoretical alternative to pattern-recognition approaches to perception. In these frameworks, perceptual systems seek to provide an explanation for sensory data using generative models. These models evaluate hypotheses about the state of the world according to how well they could give rise to (or generate) the observed properties of the perceptual input, with the credibility of hypotheses adjusted so as to minimize prediction error, that is, the deviation between the predicted and observed signal (Friston, 2005; Hinton, 2007; Rao & Ballard, 1999). The generative models themselves can be updated (and learned) in much the same way, based on the differences between the expected (i.e., predicted) signal, and the observed signal.

From the perspective of a generative model approach to spoken language processing, it would be surprising if the spoken-word recognition system did *not* use anticipatory coarticulation to generate and evaluate hypotheses about possible upcoming sounds and words, along with hypotheses about how these sounds and words would be realized in the upcoming signal. Within a data-explanation framework, pre word-onset coarticulation creates a prediction error—for instance, in the context of the materials used in the current study, a deviation from the expected acoustic properties of the vowel in the determiner (in

the absence of a hypothesis about the following word). If the particular prediction error can be accounted for by a specific hypothesis or hypotheses about the phonetic properties of upcoming segments, a generative model naturally predicts that listeners anticipate the upcoming word. Thus, the behavioral effects that we observed would arise with the current materials because there is acoustic–phonetic information in the vowel that is better predicted by the hypothesis of the determiner plus a particular upcoming segment or type of segment, rather than just the determiner alone. Our classifier analysis establishes that such information is, in fact, present early in the determiner.

Although prediction-based generative model approaches are often formalized within a framework of hierarchical Bayesian inference, they are also compatible with connectionist-based models in which differences between predicted states and the input result in an error signal that is minimized during learning (e.g., Chang, Dell, & Bock, 2006; Elman, 1990; Gaskell, 2003). We note that Jordan and Rumelhart's (1992) foundational work using forward models—a particular type of generative model— for feedback in motor control was initially formulated within such a framework. Likewise, McMurray and Jongman (2011) present a (non-generative) discriminative classification model of fricative categorization which achieves higher accuracy when using prediction error rather than raw acoustic cue values, with patterns of accuracy closely matching those of human observers. That work suggests that much of the within-category variability in cue values can be "explained away" by considering contextual variables such as neighboring vowels and talker identity. In a generative modeling framework, this corresponds to evaluating hypothesized categories based on how well they account for the prediction error which remains after accounting for preceding context (either linguistic or indexical). McMurray and Jongman's (2011) work thus complements our finding that the variability that *remains* after taking into account the current segment is informative about upcoming segments, and that listeners make immediate use of that information.

It is important to acknowledge that pattern recognition, exemplar and generative models can all, in principle, explain our results. Nonetheless, we believe that explanation-based approaches provide the most promising account. The reason is that these models not only predict immediate use of anticipatory coarticulation, as was observed in the present study, but also flexible use of this information across speakers and across linguistic contexts. In particular, the same information in the signal should receive different explanations depending upon context. Coarticulation is ubiquitous and it can be affected by a variety of different factors, including variables that affect utterance planning, speech rate, and most crucially both preceding and upcoming information. Therefore it provides an empirical domain for investigating the prediction that particular acoustic information will receive different explanations in different contexts. In general, pattern recognition approaches to coarticulation, including exemplar approaches, become less plausible if particular acoustic information receives different interpretations depending upon the contextual support for alternative explanations.

Emerging evidence suggests that lexical processing exhibits the type of flexibility that is predicted by data explanation models. For example, a syllable of relatively long duration which is phonemically consistent with the onset of a longer word, for instance, *ham* in *hamster*, biases listeners to predict that the syllable corresponds to the word "ham" (Salverda, Dahan, & McQueen, 2003). An upcoming prosodic boundary is typically marked by segmental lengthening, and the lengthened sequence /hæm/ is therefore more consistent with a short word than with the onset of a longer word—whose initial syllable would typically not be followed by a prosodic boundary. Brown, Salverda, Gunlogson, and Tanenhaus (in press) demonstrate that information structure, in particular whether a word is

discourse given or discourse new, modulates the interpretation of segmental lengthening as a cue to a prosodic boundary, as predicted by data explanation models.

Perhaps most strikingly, manipulations of "distal prosody" within an utterance, including patterns of pitch accents and stress alternation, have strong effects on how subsequent acoustical material is interpreted (Dilley & McAuley, 2008). Moreover, manipulating the speech rate of the part of an utterance that precedes a particular acoustic sequence can make words appear and disappear (e.g., hearing "are best" vs. "are our best"; Dilley & Pitt, 2010). Extensions of this work by Brown and her colleagues demonstrate that distal prosody creates on-line expectations (e.g., Brown, Dilley, & Tanenhaus, 2012; Brown, Salverda, Dilley, & Tanenhaus, 2011). These results suggest that listeners use contextual information in the speech signal to rapidly adjust their interpretation of speech, as predicted by generative models (cf. Brown et al., 2012; Farmer, Brown, & Tanenhaus, 2013; Kleinschmidt & Jaeger, 2011, 2012). The generative-model framework predicts that under appropriate circumstances, distal prosody should modulate how cues associated with coarticulation are interpreted. This is a promising direction for future research.

Finally, the fact that we found such early effects of processing the vowel schwa on the recognition of a following word is interesting from an information theoretical perspective on the transmission of speech. The English definite and indefinite articles (*the* and *a*, respectively) are short and typically have a reduced vowel. This makes them strongly subject to coarticulation, and thus ideal carriers of information pertaining to adjacent speech sounds and words. Thus, the relatively "weak" phonetic properties of many function words could be considered a functional design property of English, insofar as they facilitate the processing of neighboring speech sounds, and the recognition of following content words in continuous speech.

Acknowledgments

This research was supported by NIH Grants DC005071 and HD073890 to MKT. We thank Meredith Brown and Laura Dilley for labeling word onsets, Judith Degen for ggplot magic, and Michael Berger, Sarah Brown-Schmidt, Robert Jacobs, Florian Jaeger, Joyce McDonough, and Bob McMurray for helpful comments on this research.

Appendix A

Experimental stimulus sets

Pictures associated with each item were selected from the standardized set of pictures developed by Snodgrass and Vanderwart (1980).

Potential targets Approximant-initial	**Fricative-initial**	**Distractor**	**Distractor**
ladder	fish	grapes	pineapple
lamp	shirt	mountain	bus
leaf	foot	bottle	cannon
lemon	fly	pot	garbage can
lion	shoe	comb	pen
lobster	flag	pipe	bed

Potential targets Approximant-initial	Fricative-initial	Distractor	Distractor
rabbit	fork	peanut	kite
raccoon	scissors	corn	pumpkin
rhino	fence	basket	cake
ring	frog	bread	door
rocking chair	snake	piano	glove
ruler	sandwich	pepper	kettle
watch	sock	giraffe	carrot
well	snowman	cow	cup
wheel	spider	clothespin	gun
whistle	spoon	chair	desk
windmill	football	chain	glasses
window	flower	tie	camel

Appendix B

Statistical power in Experiment 1

Here we address the concern that we might not have observed effects of the speech signal on eye-movement control prior to 200 ms in Experiment 1 because of insufficient statistical power. We computed the effect size for the earliest effect of speech on eye-movement control in Experiment 2. To be conservative, we computed the effect size for the 110–150 ms time bin in Experiment 2, where we found a marginally significant effect. (Note that if we instead consider the first time bin in Experiment 2 with a significant effect, 130–170 ms, we obtain a larger effect size, which would result in higher values for achieved statistical power in Experiment 1 than those reported below.) The effect size in that time bin is $w = .56$. At the suggestion of a reviewer, we also considered the effect size in the 80– 120 ms bin in Altmann's (2011) reanalysis of the Kamide et al. (2003) data, which one could argue to be a better a priori estimate of the effect size of the earliest effects of speech on eye-movement control. On the basis of the data reported in Fig. 4 and Table 1 in Altmann (2011), we estimate the effect size in the 80–120 ms bin to be $w = .55$ (assuming 36 trials with a saccade to the target, and 21 trials with a saccade to the associated distractor).

We believe that the relevant effect size for the computation of statistical power in pre 200-ms time bins in Experiment 1 corresponds to that obtained in our Experiment 2 rather than the (estimated) effect size for Altmann's (2011) reanalysis of the Kamide et al. (2003) data. First, both of our experiments (but not the experiments reanalyzed by Altmann) had been designed to detect early effects of the speech signal on eye-movement control. Second, there are differences in stimuli, design and procedure between our studies and the Kamide et al. study that may have influenced the effect size obtained for the latter study. However, we note that the effect size in the earliest bin showing a significant effect for the Kamide et al. data reported in Altmann (2011), $w = .55$, is virtually identical to that obtained in the earliest bin showing a marginal effect in our Experiment 2, $w = .56$. Therefore, the statistical power computations reported below would give very similar results if they were performed using the effect size obtained in Altmann's study.

We performed a power analysis on the data from Experiment 1, using the freely available computer program G*Power 3.1.7 (Faul, Erdfelder, Lang, & Buchner, 2007). For each 40-

ms analysis time bin, starting with the 0– 40 ms bin and ending with the 200–240 bin, using a 10-ms moving window, we computed the power of this experiment to detect an effect of $w = .56$, assuming $\alpha = .05$, and given the sample size (that is, the number of trials with a saccade to the target plus the number of trials with a saccade to the associated distractor). The results are plotted in Fig. 13. With the exception of the very early 30–70 ms and 40–80 ms time bins, all 40-ms time bins prior to 200 ms included enough trials to detect an effect of $w = .56$ at well over 90% power. On the basis of this analysis, we conclude that it is unlikely that: (a) information in the speech signal influences eye-movement control earlier than 200 ms in Experiment 1; (b) this effect was of a magnitude as large or larger than the earliest effect observed in Experiment 2; and (c) the effect was not detected due to a lack of statistical power.

References

Allopenna P, Magnuson JS, Tanenhaus MK. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. Journal of Memory and Language. 1998; 38:419–439.

Altmann GTM. Language-mediated eye movements in the absence of a visual world: The “blank screen paradigm”. Cognition. 2004; 93:B79–B87. [PubMed: 15147941]

Altmann GTM. Language can mediate eye movement control within 100 milliseconds, regardless of whether there is anything to move the eyes to. Acta Psychologica. 2011; 137:190–200. [PubMed: 20965479]

Beddor PS, McGowan KB, Boland JE, Coetzee AW, Brasher A. The time course of perception of coarticulation. Journal of the Acoustical Society of America. 2013; 133:2350–2366. [PubMed: 23556601]

Boersma P, Weenink D. Praat: Doing phonetics by computer [Computer program]. Version 5.3.35. 2012 <http://www.praat.org/> Retrieved 08.12.12.

Brown, M.; Dilley, LC.; Tanenhaus, MK. Real-time expectations based on context speech rate can cause words to appear or disappear. In: Miyake, N.; Peebles, D.; Cooper, RP., editors. Proceedings of the 34th annual conference of the cognitive science society. Austin, TX: Cognitive Science Society; 2012. p. 1374-1379.

Brown M, Salverda AP, Gunlogson C, Tanenhaus MK. Interpreting prosodic cues in discourse context. Language, Cognition and Neuroscience. 2013 (in press).

Brown M, Salverda AP, Dilley LC, Tanenhaus MK. Expectations from preceding prosody influence segmentation in online sentence processing. Psychonomic Bulletin & Review. 2011; 18:1189–1196. [PubMed: 21968925]

Chang F, Dell GS, Bock K. Becoming syntactic. Psychological Review. 2006; 113:234–272. [PubMed: 16637761]

Coenen E, Zwitserlood P, Bölte J. Variation and assimilation in German: Consequences for lexical access and representation. Language and Cognitive Processes. 2001; 16:535–564.

Cole J, Linebaugh G, Munson CM, McMurray B. Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. Journal of Phonetics. 2010; 38:167–184. [PubMed: 21173864]

Cooper RM. The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. Cognitive Psychology. 1974; 6:84–107.

Dahan D, Magnuson JS, Tanenhaus MK, Hogan E. Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. Language and Cognitive Processes. 2001; 16:507–534.

Dahan D, Tanenhaus MK. Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2004; 30:498–513.

Dilley LC, McAuley JD. Distal prosodic context affects word segmentation and lexical processing. Journal of Memory and Language. 2008; 59:294–311.

Dilley LC, Pitt MA. Altering context speech rate can cause words to appear or disappear. Psychological Science. 2010; 21:1664–1670. [PubMed: 20876883]

Elman JL. Finding structure in time. Cognitive Science. 1990; 14:179–211.

Elman, JL.; McClelland, JL. Exploiting lawful variability in the speech wave. In: Perkell, JS.; Klatt, DH., editors. Invariance and variability in speech processes. Hillsdale, NJ: Erlbaum; 1986. p. 360-384.

Farmer TA, Brown M, Tanenhaus MK. Prediction, explanation, and the role of generative models in language processing. Behavioral and Brain Sciences. 2013; 36:211–212. [PubMed: 23663410]

Faul F, Erdfelder E, Lang A-G, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods. 2007; 39:175–191. [PubMed: 17695343]

Fowler CA. Production and perception of coarticulation among stressed and unstressed vowels. Journal of Speech and Hearing. 1981; 24:127–139.

Fowler CA. Parsing coarticulated speech in perception: Effects of coarticulation resistance. Journal of Phonetics. 2005; 33:199–213.

Friston K. A theory of cortical responses. Philosophical transactions of the Royal Society of London. Series B, Biological sciences. 2005; 360:815–836.

Gaskell MG. Modelling regressive and progressive effects of assimilation in speech perception. Journal of Phonetics. 2003; 31:447–463.

Gaskell MG, Marslen-Wilson WD. Phonological variation and inference in lexical access. Journal of Experimental Psychology: Human Perception and Performance. 1996; 22:144–158. [PubMed: 8742258]

Gaskell MG, Marslen-Wilson WD. Lexical ambiguity resolution and spoken word recognition: Bridging the gap. Journal of Memory and Language. 2001; 44:325–349.

Gaskell MG, Snoeren ND. The impact of strong assimilation on the perception of connected speech. Journal of Experimental Psychology: Human Perception and Performance. 2008; 34:1632–1647. [PubMed: 19045997]

Goldinger SD. Echoes of echoes? An episodic theory of lexical access. Psychological Review. 1998; 105:251–279. [PubMed: 9577239]

Gow DW Jr. Does English coronal place assimilation create lexical ambiguity? Journal of Experimental Psychology: Human Perception and Performance. 2002; 28:163–179.

Gow DW Jr. Feature parsing: Feature cue mapping in spoken word recognition. Perception & Psychophysics. 2003; 65:575–590. [PubMed: 12812280]

Gow, DW., Jr; McMurray, B. Word recognition and phonology: The case of English coronal place assimilation. In: Cole, JS.; Hualdo, J., editors. Papers in laboratory phonology. Vol. Vol. 9. New York: Mouton de Gruyter; 2007. p. 173-200.

Hallett, PE. Eye movements. In: Boff, KR.; Kaufman, L.; Thomas, JP., editors. Handbook of perception and human performance. New York, NY: Wiley; 1986. p. 10.1-10.112.

Hardcastle, WJ.; Hewlett, N. Coarticulation: Theory, data and techniques. Cambridge: Cambridge University Press; 1999.

Hinton GE. Learning multiple layers of representation. Trends in Cognitive Sciences. 2007; 11:428–434. [PubMed: 17921042]

Johnson, K. Speech perception without speaker normalization: An exemplar model. In: Johnson; Mullennix, editors. Talker variability in speech processing. San Diego: Academic Press; 1997. p. 145-165.

Jordan MI, Rumelhart DE. Forward models: Supervised learning with a distal teacher. Cognitive Science. 1992; 16:307–354.

Kamide Y, Altmann GTM, Haywood SL. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. Journal of Memory and Language. 2003; 49:133–156.

Kleinschmidt, DF.; Jaeger, TF. Proceedings of the 2 nd annual workshop on cognitive modeling and computational linguistics (CMCL). Portland, OR: Association of Computational Linguistics; 2011. A Bayesian belief updating model of phonetic recalibration and selective adaptation; p. 10-19.

Kleinschmidt, DF.; Jaeger, TF. A continuum of phonetic adaptation: Evaluating an incremental belief-updating model of recalibration and selective adaptation. In: Miyake, N.; Peebles, D.; Cooper, RP., editors. Proceedings of the 34 th annual conference of the cognitive science society. Austin, TX: Cognitive Science Society; 2012. p. 605-610.

Kuehn DP, Moll KL. Perceptual effects of forward articulation. Journal of Speech and Hearing Research. 1972; 15:654–664. [PubMed: 5080057]

Magen HS. The extent of vowel-to-vowel coarticulation in English. Journal of Phonetics. 1997; 25:187–205.

Marslen-Wilson WD. Functional parallelism in spoken wordrecognition. Cognition. 1987; 25:71–102. [PubMed: 3581730]

Marslen-Wilson W, Warren P. Levels of perceptual representation and process in lexical access. Psychological Review. 1994; 101:653–675. [PubMed: 7984710]

Martin JG, Bunnell HT. Perception of anticipatory coarticulation effects. Journal of the Acoustical Society of America. 1981; 69:559–567. [PubMed: 7462478]

Matin E, Shao KC, Boff KR. Saccadic overhead: Information-processing time with and without saccades. Perception & Psychophysics. 1993; 53:372–380. [PubMed: 8483701]

McClelland JL, Elman JL. The TRACE model of speech perception. Cognitive Psychology. 1986; 18:2–85.

McMurray B, Jongman A. What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. Psychological Review. 2011; 118:219–246. [PubMed: 21417542]

McQueen JM, Norris D, Cutler A. Lexical influence in phonetic decision making: Evidence from subcategorical mismatches. Journal of Experimental Psychology: Human Perception and Performance. 1999; 25:1363–1389.

McQueen JM, Viebahn MC. Tracking recognition of spoken words by tracking looks to printed words. Quarterly Journal of Experimental Psychology. 2007; 60:661–671.

Mitterer H, Reinisch E. No delays in application of perceptual learning in speech recognition: Evidence from eye tracking. Journal of Memory and Language. 2013; 69:527–545.

Norris D. Shortlist: A connectionist model of continuous speech recognition. Cognition. 1994; 52:189–234.

Norris D, McQueen JM. Shortlist B: A Bayesian model of continuous speech recognition. Psychological Review. 2008; 115:357–395. [PubMed: 18426294]

Oller DK. The effect of position in utterance on speech segment duration in English. Journal of the Acoustical Society of America. 1973; 54:1235–1247. [PubMed: 4765808]

Pierrehumbert JB. Word-specific phonetics. Laboratory Phonology. 2002; 7:1–24.

Rao RPN, Ballard DH. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience. 1999; 2:79–87.

Salverda AP, Dahan D, McQueen JM. The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. Cognition. 2003; 90:51–89. [PubMed: 14597270]

Salverda AP, Tanenhaus MK. Tracking the time course of orthographic information in spoken-word recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2010; 36:1108–1117.

Sedivy JE, Tanenhaus MK, Chambers CG, Carlson GN. Achieving incremental interpretation through contextual representation: Evidence from the processing of adjectives. Cognition. 1999; 71:109–147. [PubMed: 10444906]

Smits R, Warner N, McQueen JM, Cutler A. Unfolding of phonetic information over time: A database of Dutch diphone perception. Journal of the Acoustical Society of America. 2003; 113:563–574. [PubMed: 12558292]

Snodgrass JG, Vanderwart M. A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. Journal of Experimental Psychology: Human Learning and Memory. 1980; 6:174–215. [PubMed: 7373248]

Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, Sedivy JC. Integration of visual and linguistic information in spoken language comprehension. Science. 1995; 268:1632–1634. [PubMed: 7777863]

Tobin SJ, Cho PW, Jennett P, Fowler C, Magnuson J. Effects of coarticulation resistance on lexical access. (in preparation).

Trottier L, Pratt J. Visual processing of targets can reduce saccadic latencies. Vision Research. 2005; 45:1349–1354. [PubMed: 15743605]

Whalen DH. Subcategorical phonetic mismatches slow phonetic judgments. Perception & Psychophysics. 1984; 35:49–64. [PubMed: 6709474]

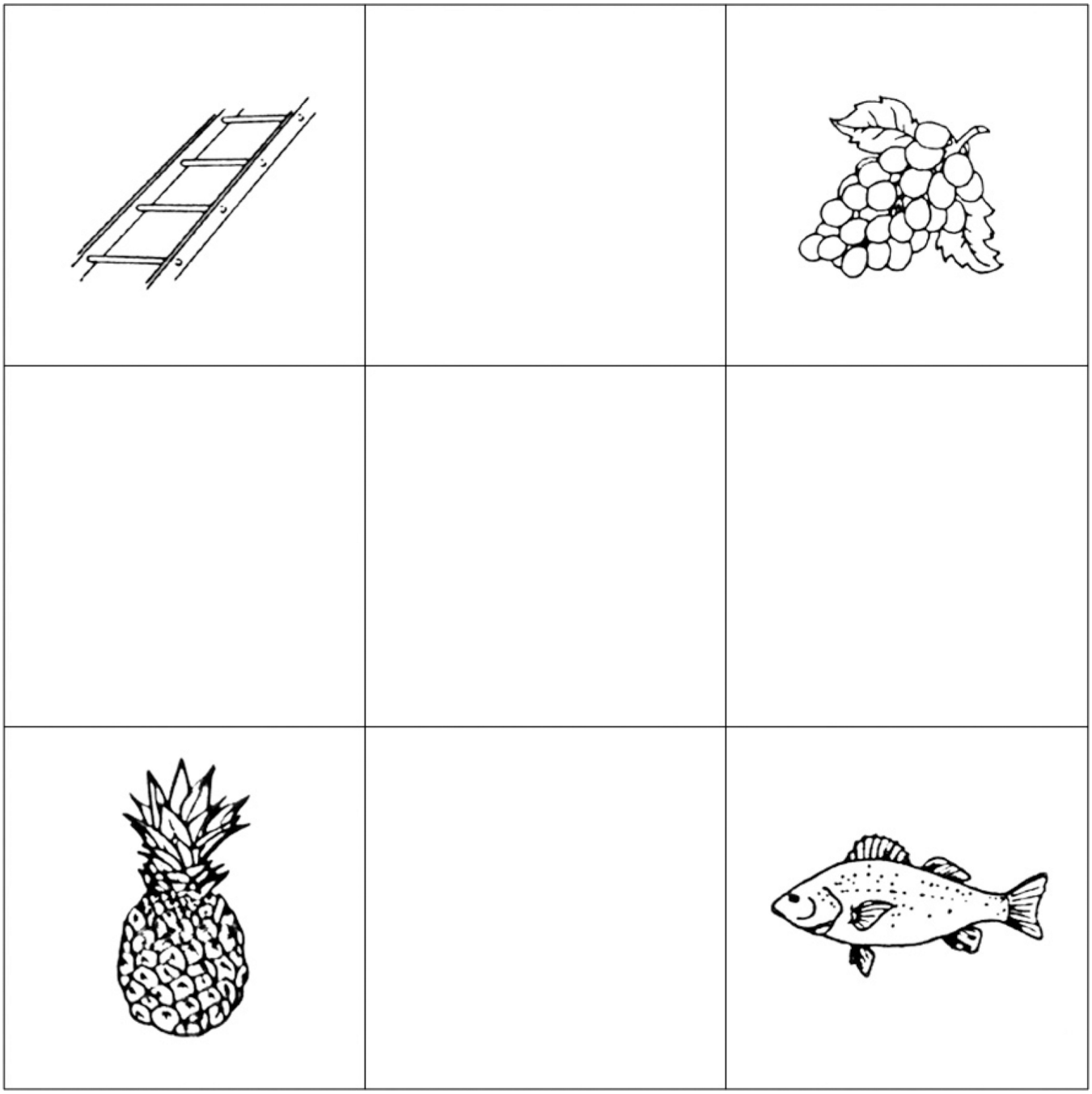

Fig. 1.
Example of a visual display used in the experiments. The ladder and the fish are potential targets; that is, for a particular participant, one of these two pictures was the target.

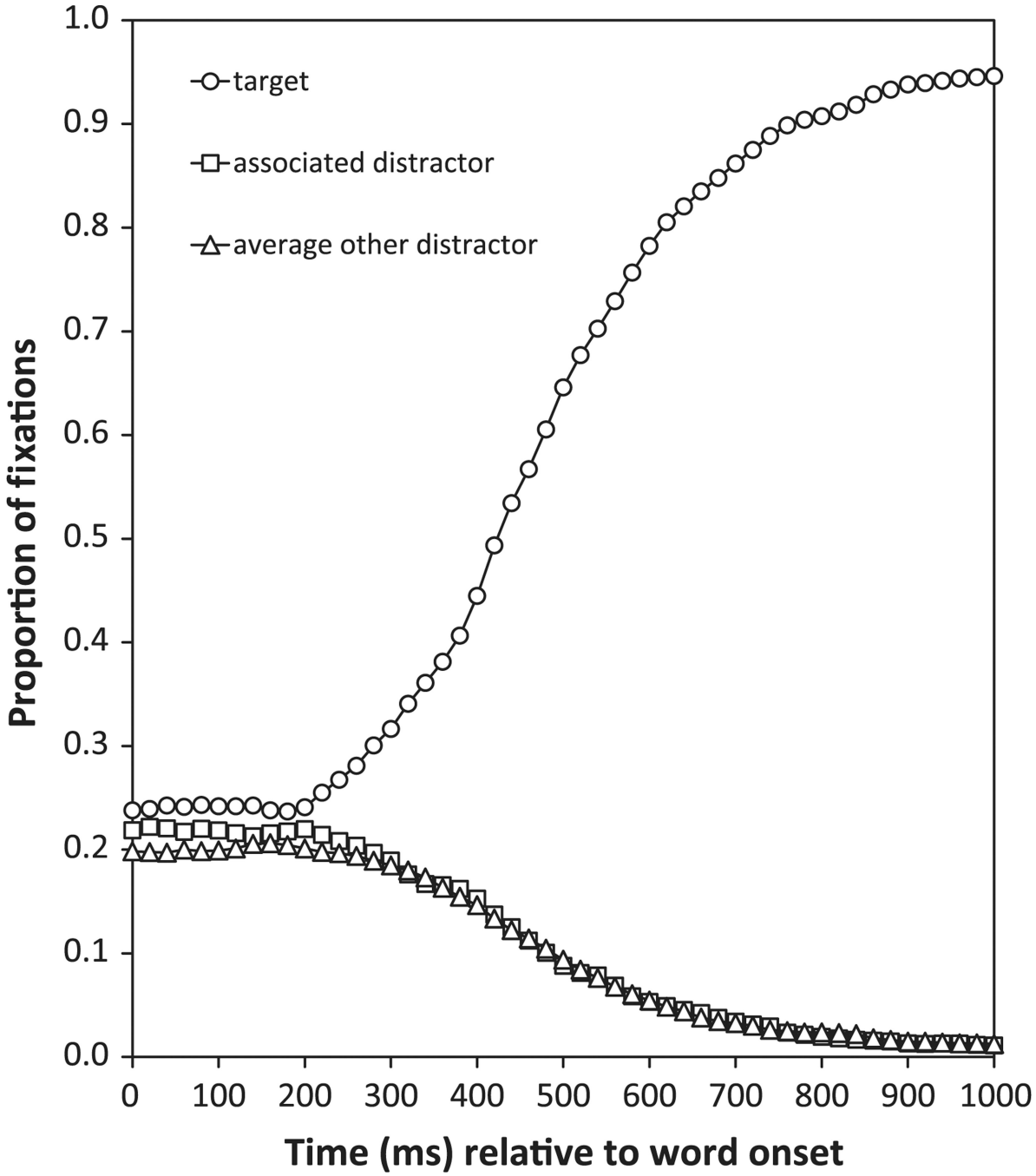

Fig. 2.
Proportion of fixations to the target, the associated distractor, and the averaged other distractors in Experiment 1, relative to the onset of the target word.

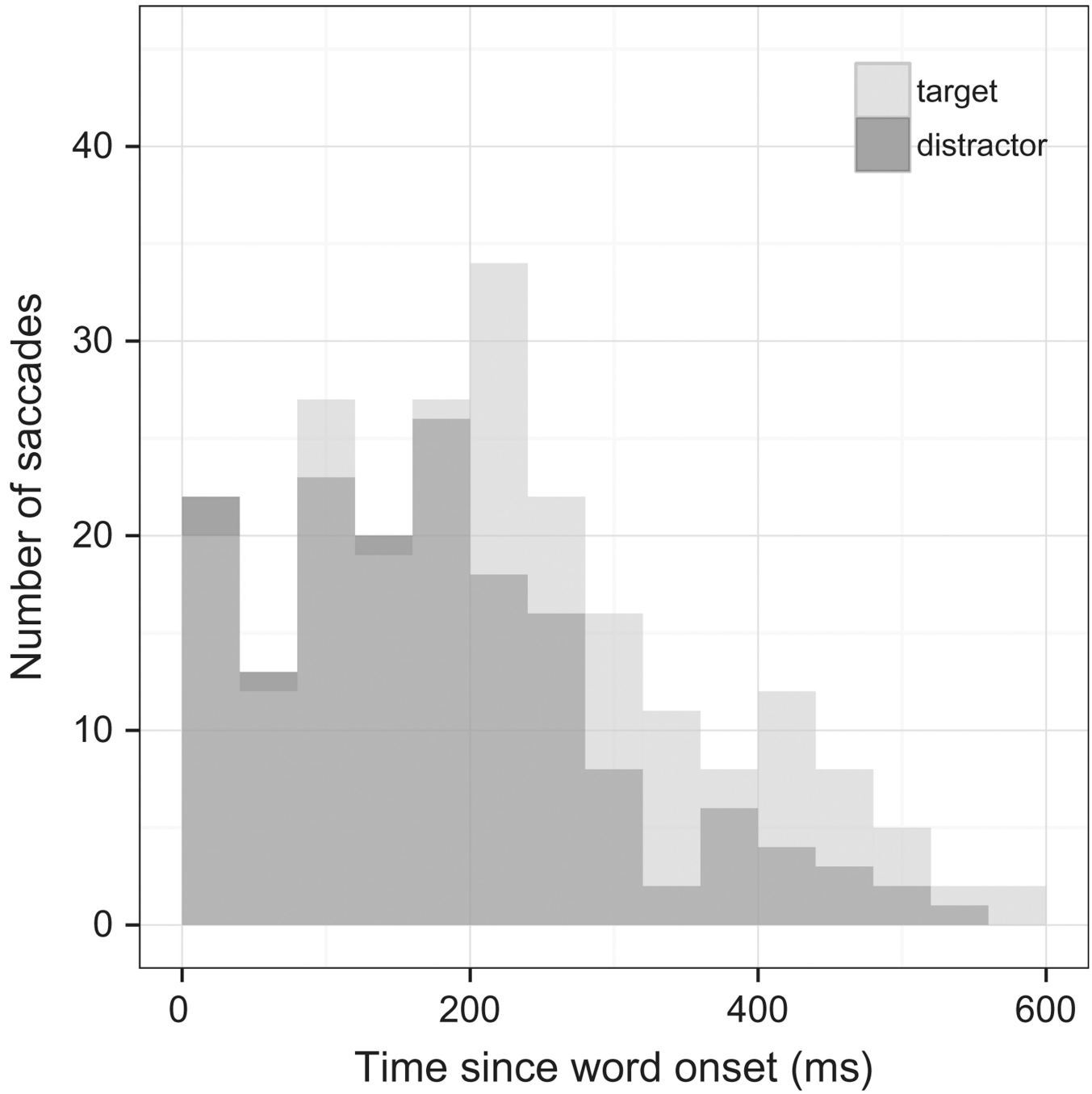

Fig. 3.
Histogram of saccade latencies for Experiment 1, in 40-ms bins. Note that the two distributions are transparent. The overlap between the distributions shows up as a mid-gray. For example, in the 0–40 ms bin there are 20 saccades to the target and 22 to the associated distractor; in the 200–240 ms bin there are 34 saccades to the target and 18 to the associated distractor.

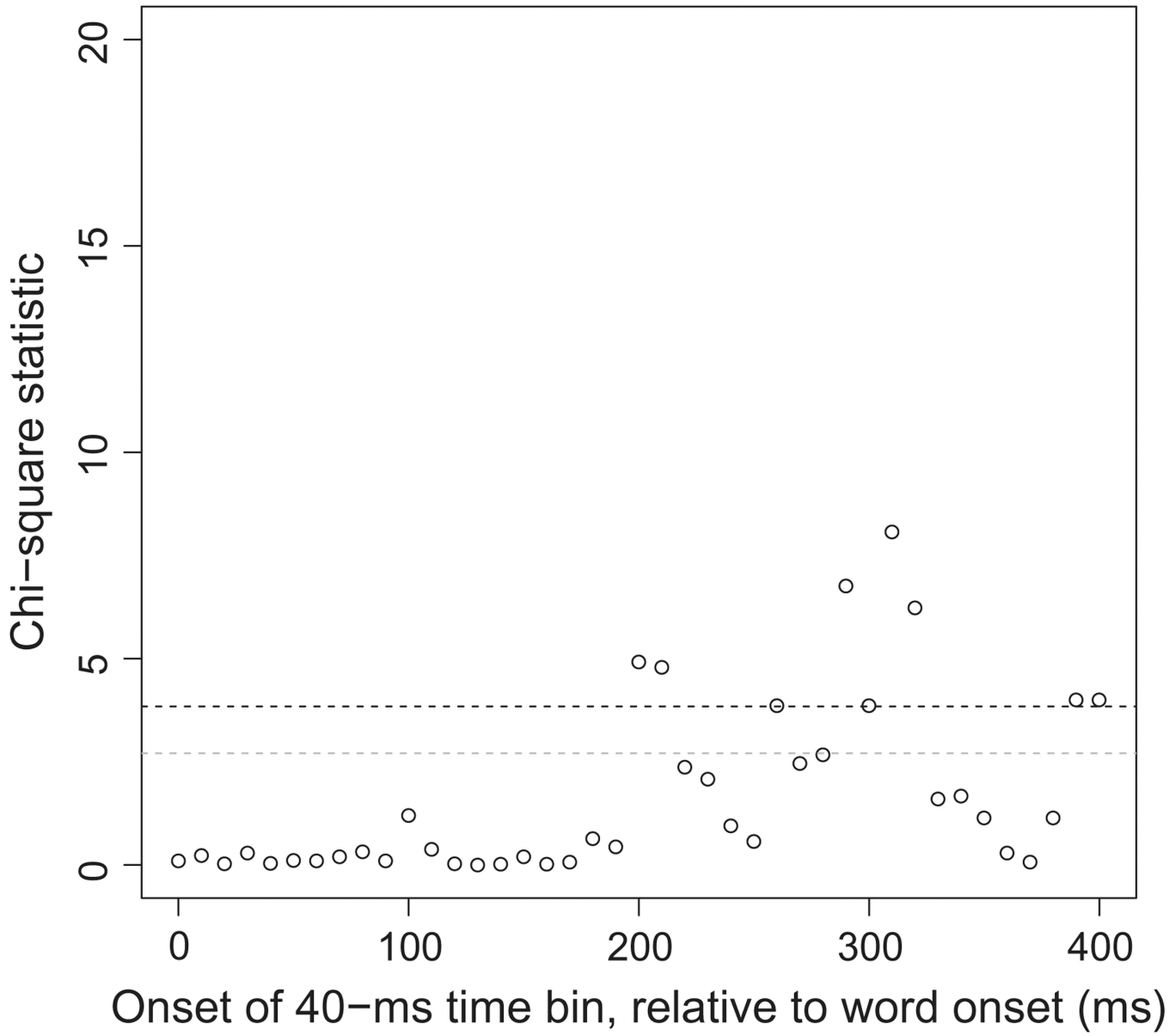

Fig. 4.

Moving-window analysis of saccade latencies in Experiment 1. Each data point represents the value of the Chi Square statistic comparing the number of trials with a saccade to the target vs. the number of trials with a saccade to the associated distractor, in a 40-ms bin relative to the onset of the target word. The black horizontal dashed line indicates the critical statistic for $p < .05$ with one degree of freedom ($\chi^2 = 3.84$); the gray horizontal dashed line indicates the statistic associated with a marginally significant effect ($\chi^2 = 2.71$).

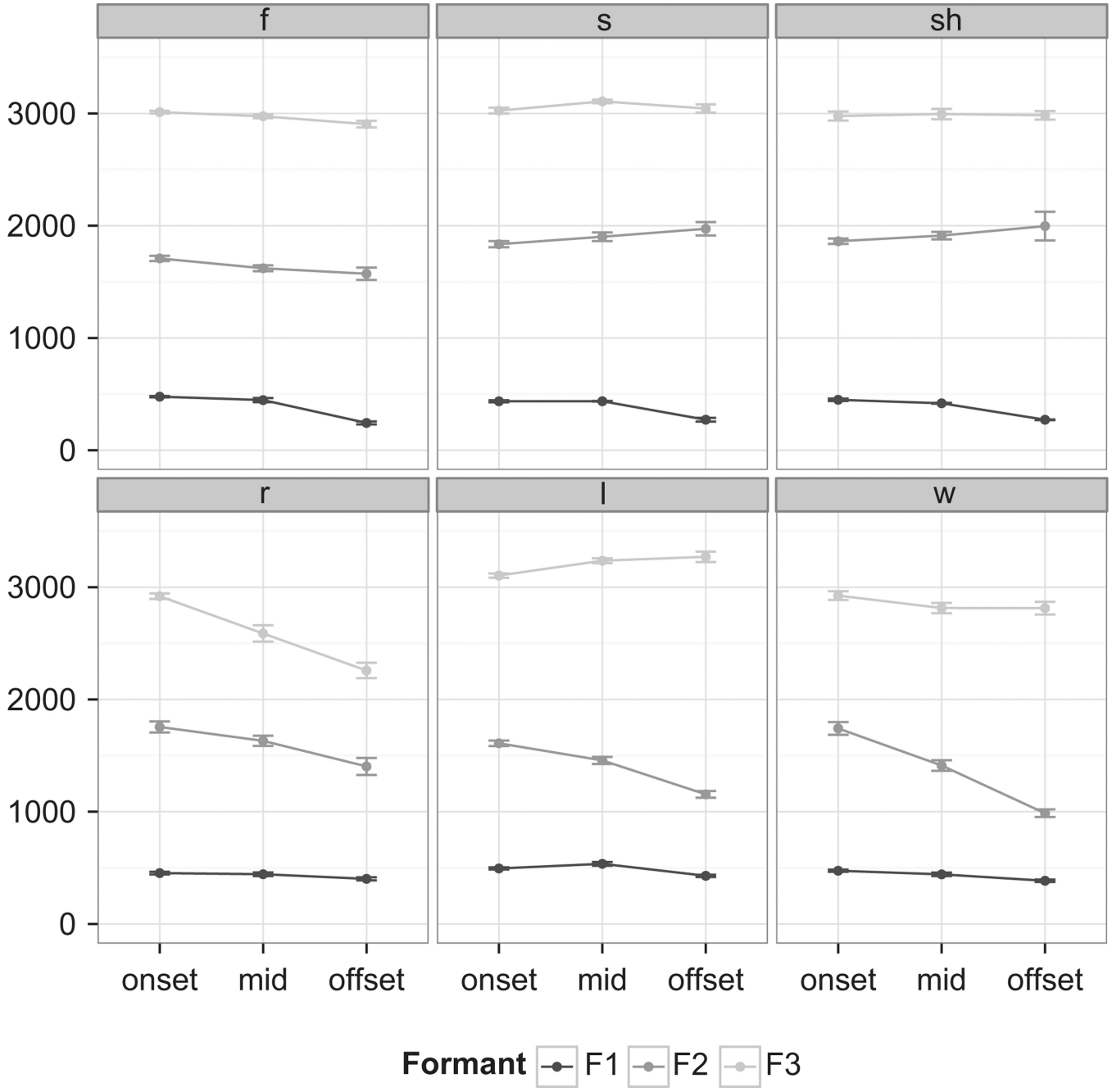

Fig. 5.

Average formant frequencies at the onset, midpoint, and offset of the vowel schwa in the determiner preceding the target word, averaged across items sharing the same initial sound. Error bars represent the standard error of the mean.

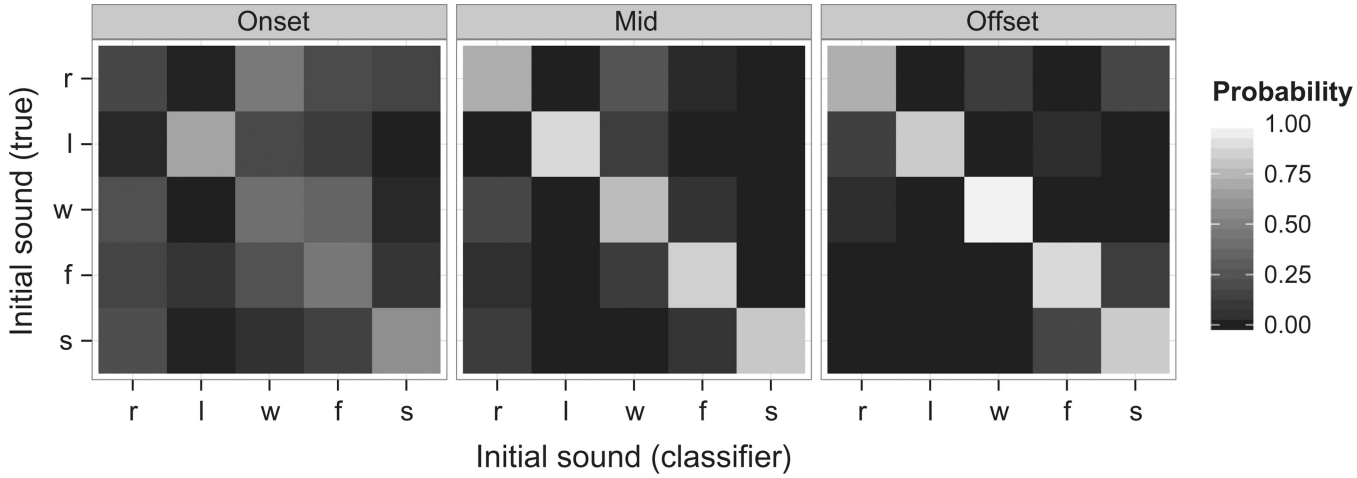

Fig. 6.

Confusion matrix for the three classifiers used to predict the initial sound of the target word on the basis of formant center frequencies at the onset, midpoint, or offset of the vowel schwa in the determiner preceding the target word.

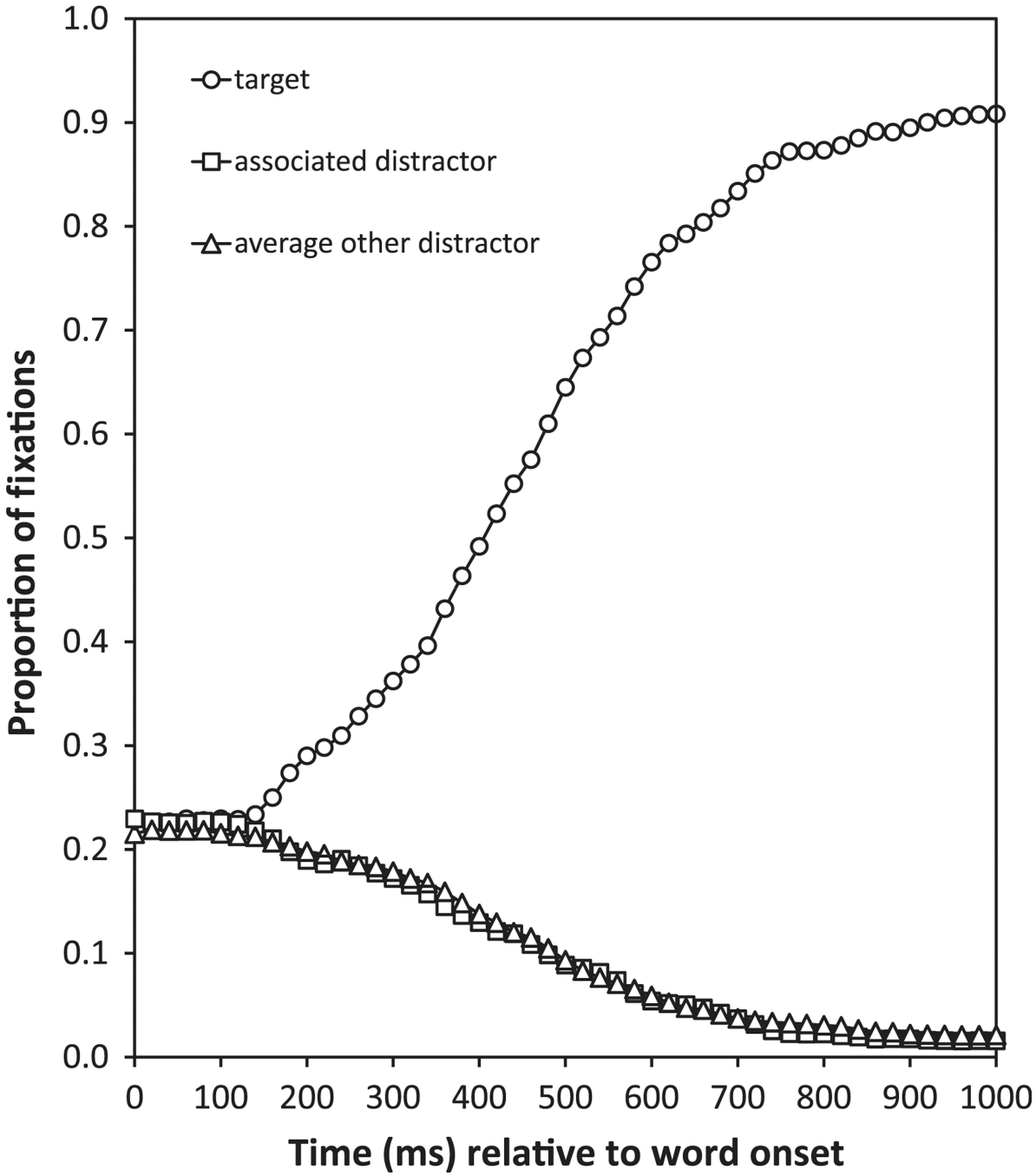

Fig. 7.
Proportion of fixations to the target, the associated distractor, and the averaged other distractors in Experiment 2, relative to the (estimated) onset of the target word.

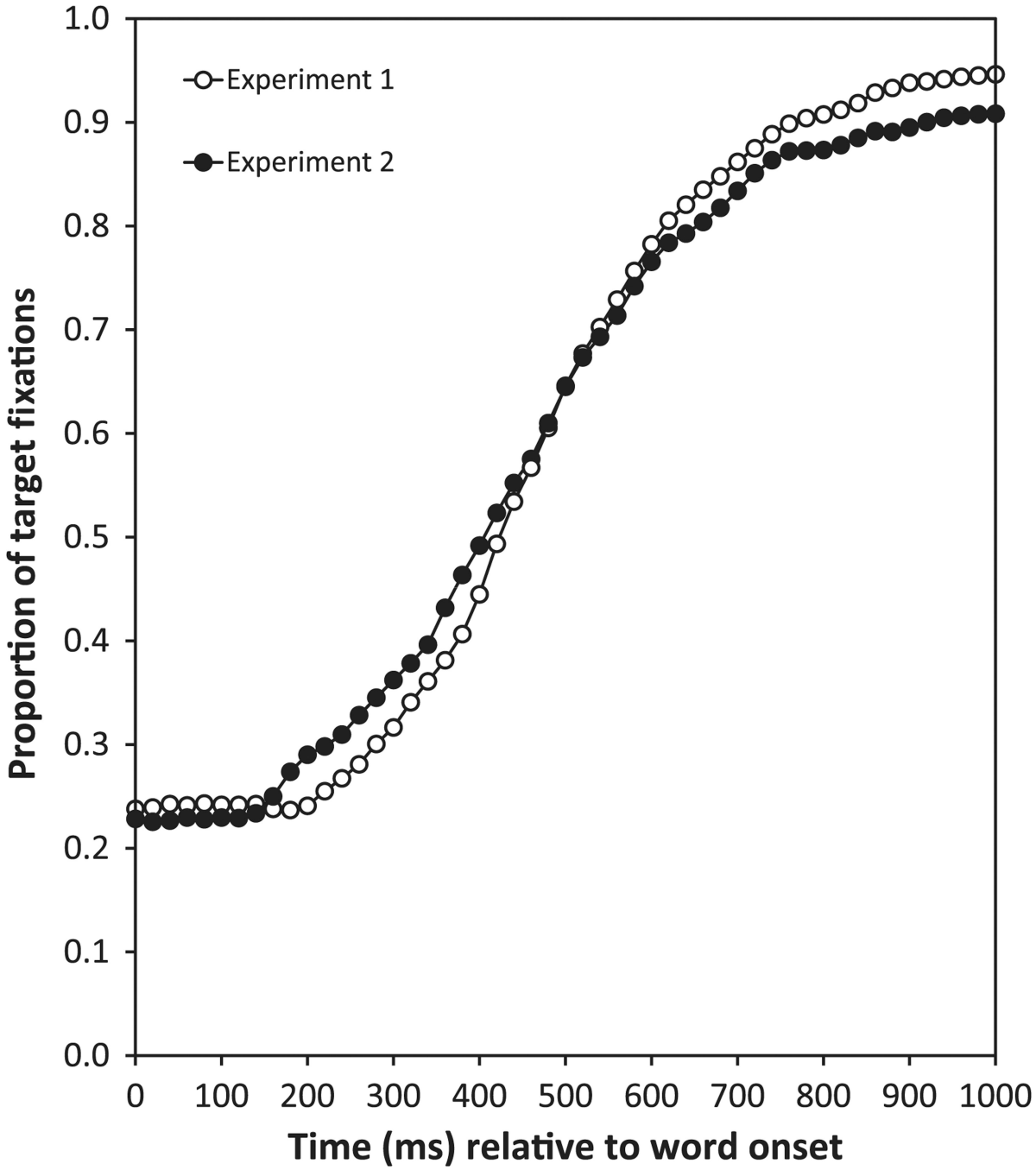

Fig. 8.
Proportion of fixations to the target in Experiment 1 and Experiment 2, relative to the onset of the target word.

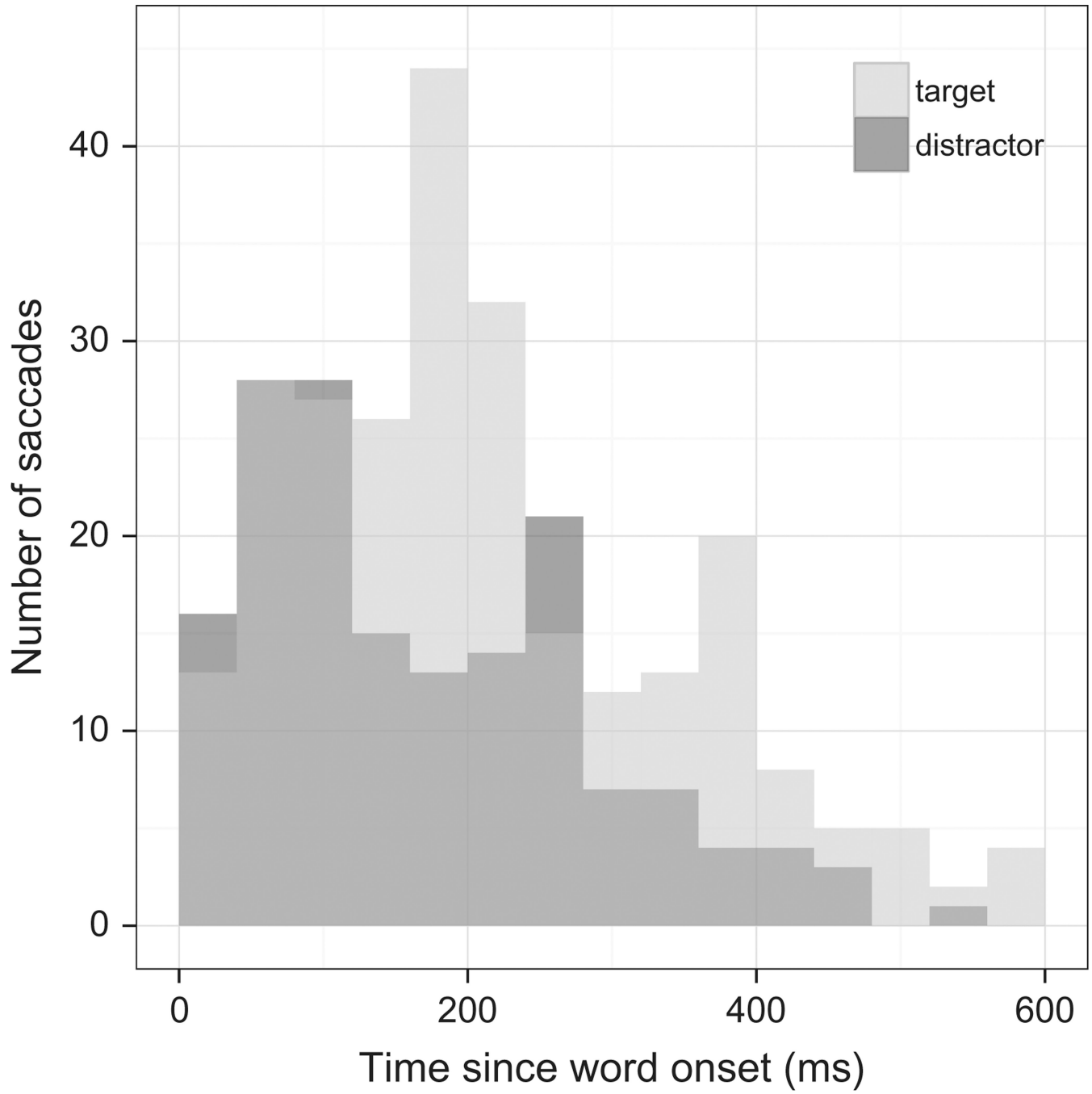

Fig. 9.
Histogram of saccade latencies for Experiment 2, in 40-ms bins.

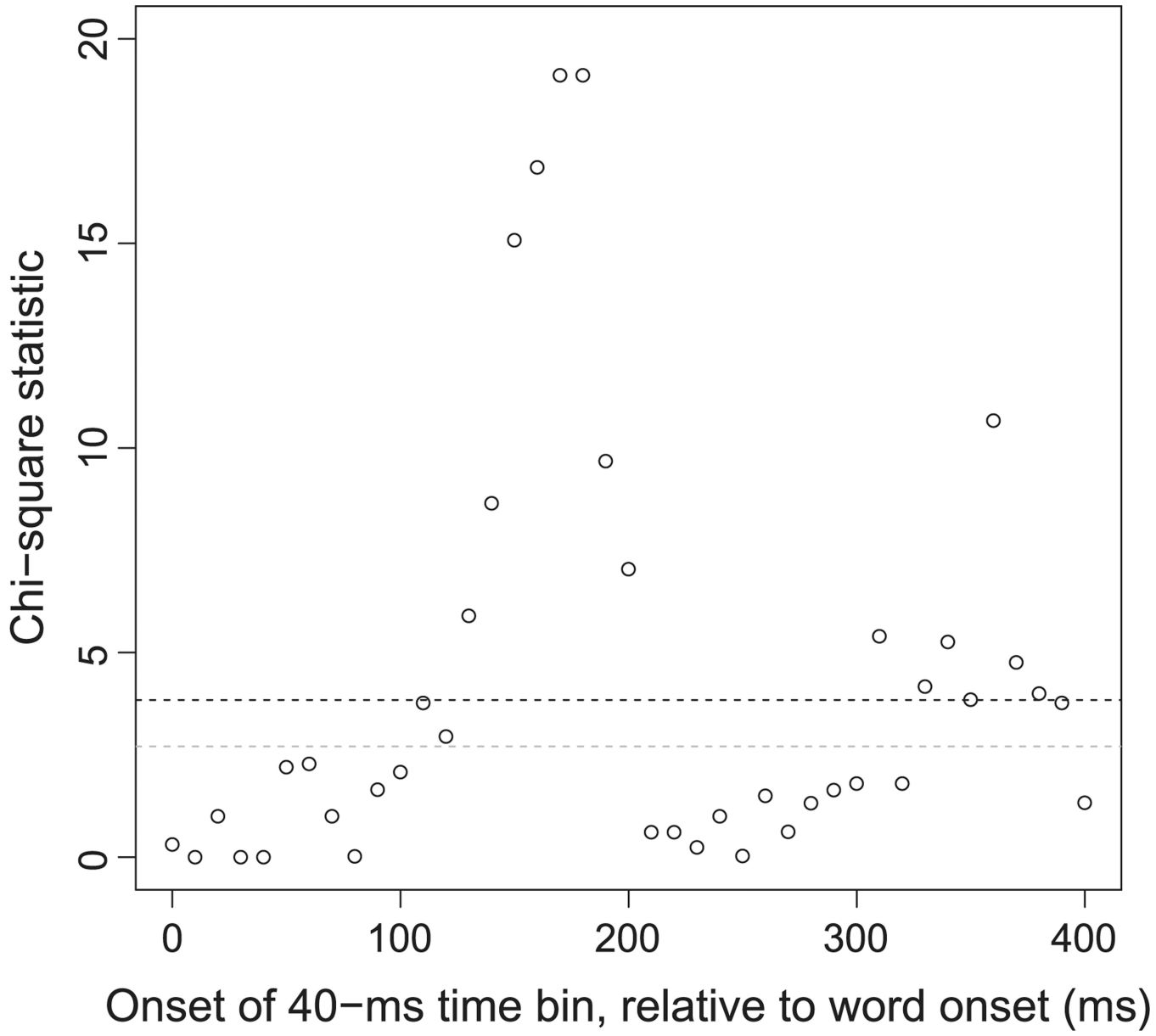

Fig. 10.

Moving-window analysis of saccade latencies in Experiment 2. Each data point represents the value of the Chi Square statistic comparing the number of trials with a saccade to the target vs. the number of trials with a saccade to the associated distractor, in a 40-ms bin, relative to the (estimated) onset of the target word. The black horizontal dashed line indicates the critical statistic for $p < .05$ with one degree of freedom ($\chi^2 = 3.84$); the gray horizontal dashed line indicates the statistic associated with a marginally significant effect ($\chi^2 = 2.71$).

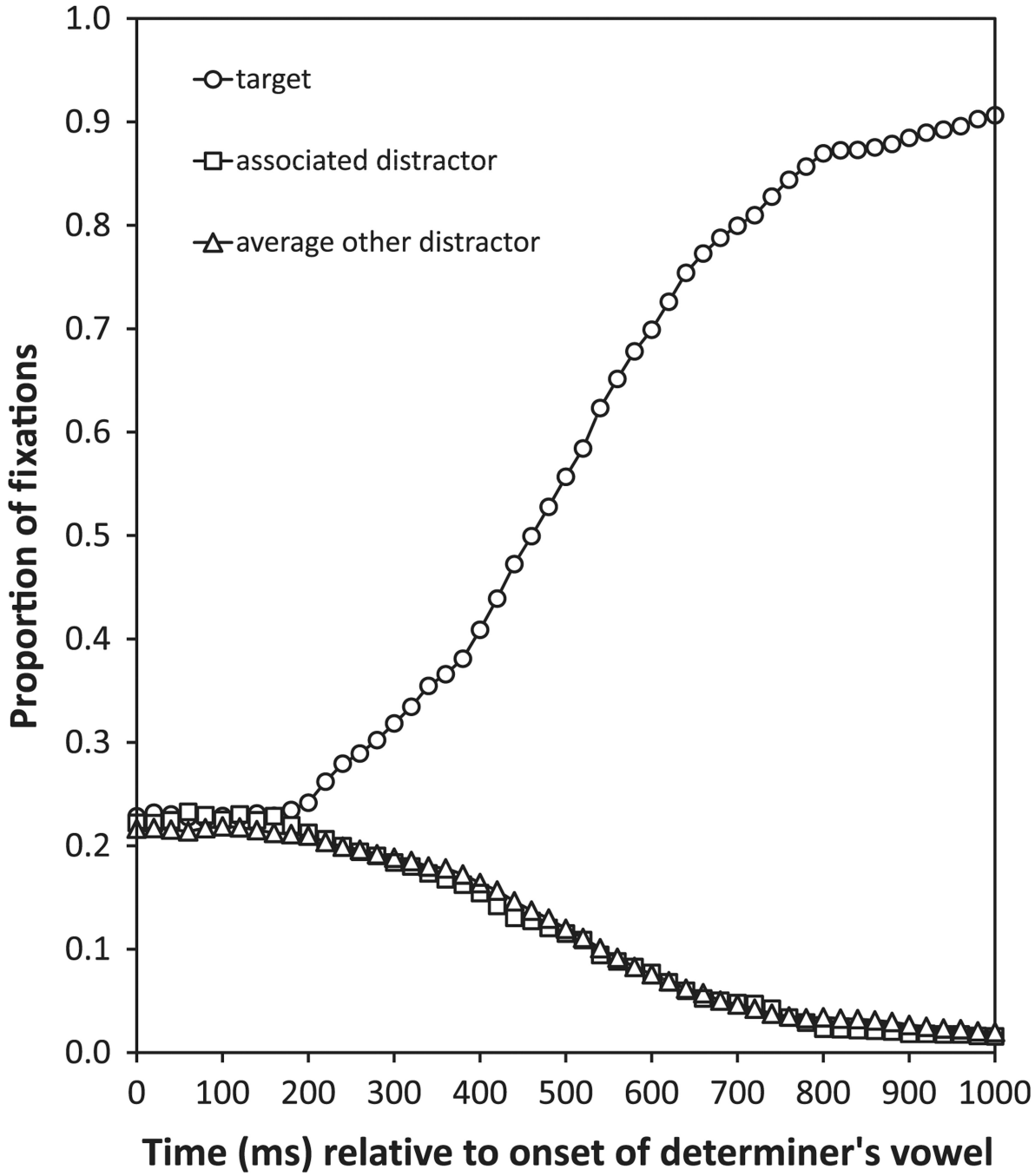

Fig. 11.
Proportion of fixations to the target, associated distractor, and averaged other distractors in Experiment 2, relative to the onset of the vowel of the determiner.

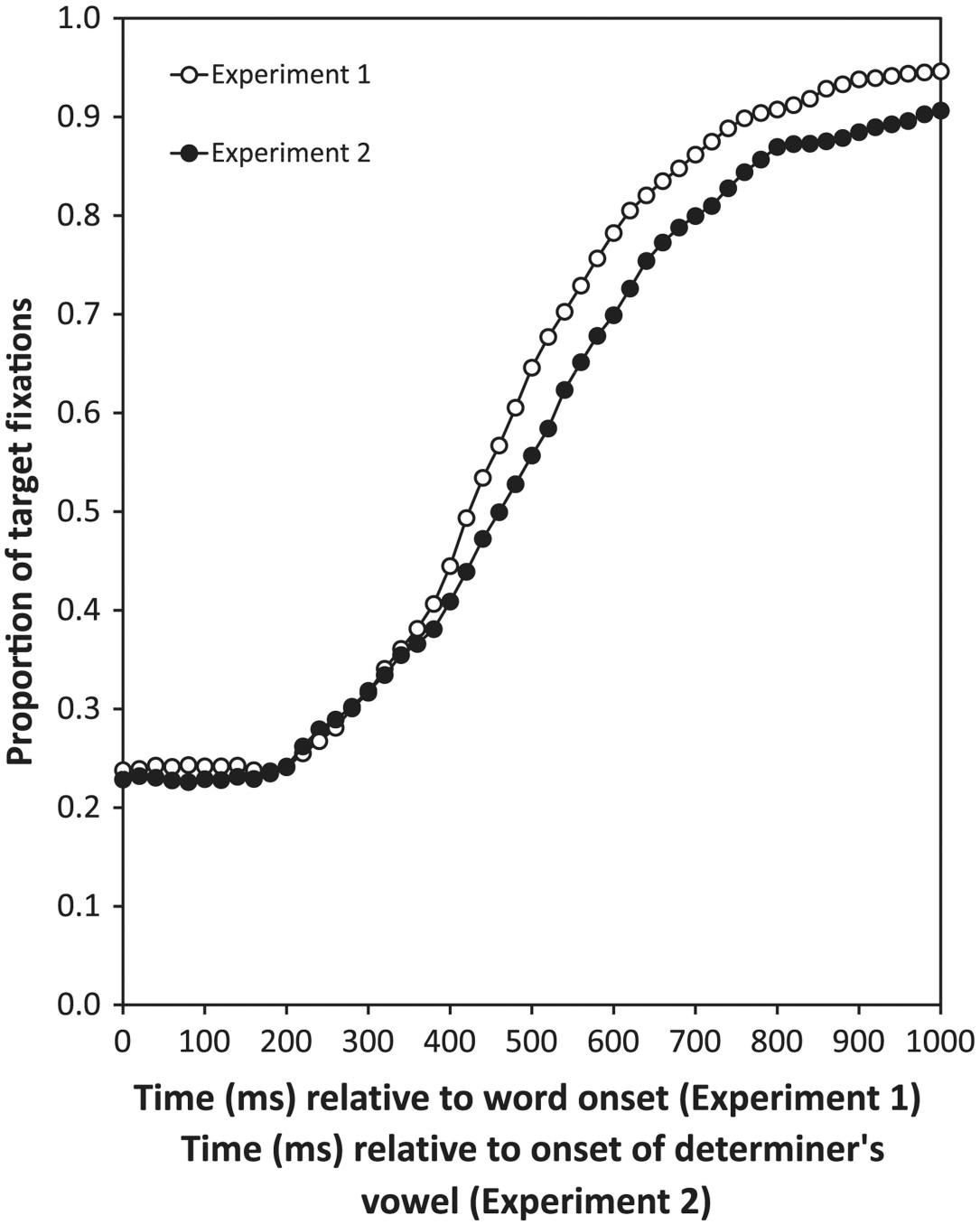

Fig. 12.
Proportion of fixations to the target in Experiment 1 (relative to the onset of the target word) and Experiment 2 (relative to the onset of the vowel of the determiner).

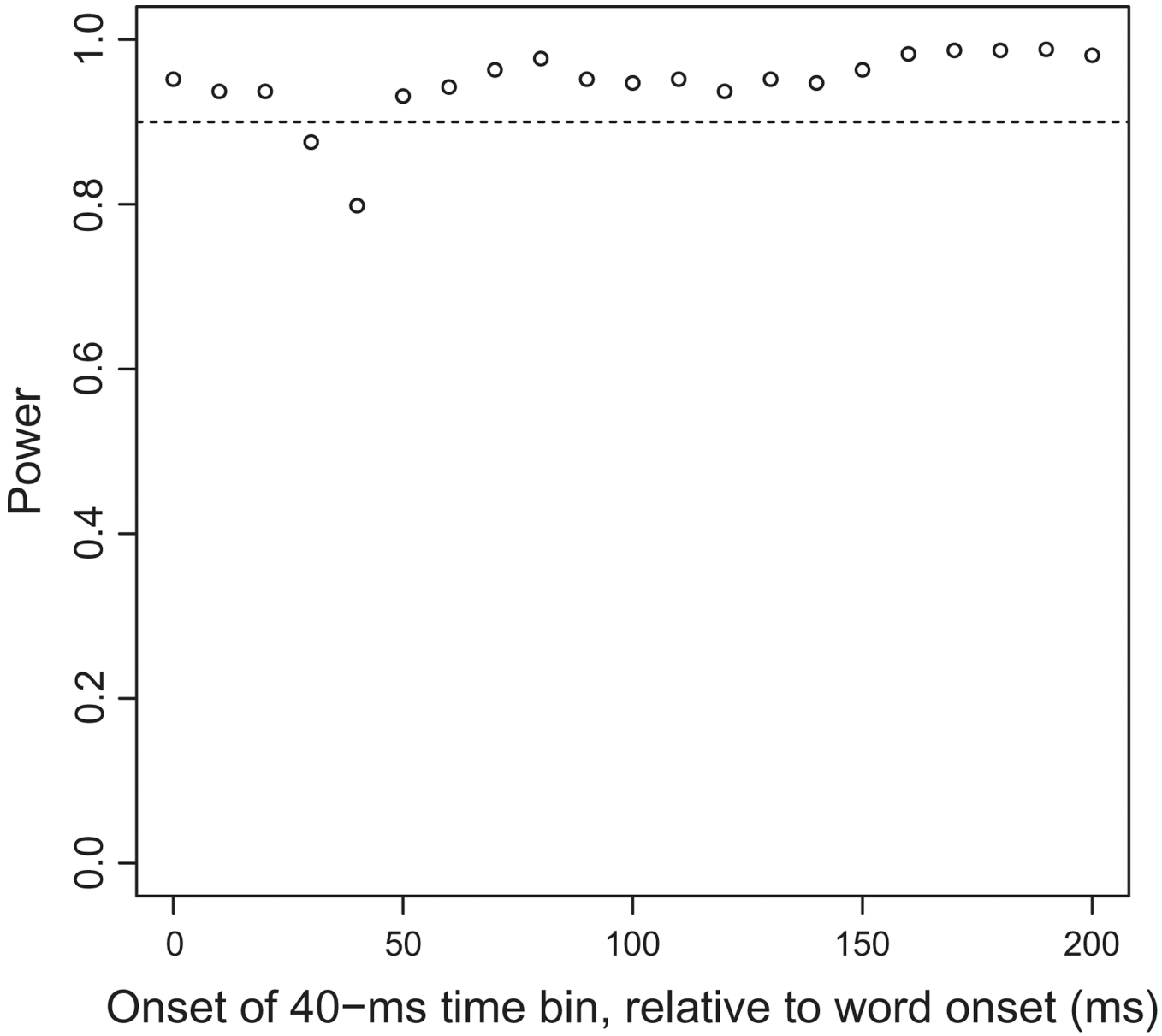

Fig. 13.
Statistical power for each of the time bins in the moving-window analysis of the data from Experiment 1 to detect an effect of $w = .56$, assuming $\alpha = .05$.