The temporal dynamics of spoken word recognition

in adverse listening conditions

Susanne Brouwer and Ann R. Bradlow

Address of correspondence

Linguistics Department

Northwestern University

2016 Sheridan Road

Evanston, IL 60208

USA

s.m.brouwer@uu.nl

Abstract

This study examined the temporal dynamics of spoken word recognition in noise and background speech. In two visual-world experiments, English participants listened to target words while looking at four pictures on the screen: a target (e.g. *candle*), an onset competitor (e.g. *candy*), a rhyme competitor (e.g. *sandal*), and an unrelated distractor (e.g. *lemon*). Target words were presented in quiet, mixed with broadband noise, or mixed with background speech. Results showed that lexical competition changes throughout the observation window as a function of what is presented in the background. These findings suggest that, rather than being strictly sequential, stream segregation and lexical competition interact during spoken word recognition.

Keywords: spoken word recognition; lexical competition; stream segregation; eye-tracking

Introduction

Successful communication depends on accurate word recognition. An extensive body of research shows that the speech signal is processed continuously involving activation of multiple lexical candidates which are concurrently considered and evaluated (see McQueen, 2007). For instance, as listeners hear the word *candle*, similar sounding lexical candidates, such as *candy* and *sandal*, will be considered in parallel and actively compete with each other until they mismatch with the input. This account of spoken word recognition presumes that competition arises from the sound-based organization of the listener's mental lexicon. However, in situations of speech recognition under adverse listening conditions, there is likely to be additional competition or interference from sounds in the background. The aim of this paper is to examine how adverse listening conditions due to the presence of background noise and background speech impact the dynamics of spoken word recognition.

A highly applicable technique for providing detailed time-course information about lexical access in continuous speech is the visual-world paradigm (Cooper, 1974; Tanenhaus, Spivey, Knowlton, Eberhard, & Sedivy, 1995). In this method, listeners' eye movements are measured as they listen to speech and see pictures of objects on a screen. Importantly, eye movements are closely time-locked to the speech input and thus reflect the degree of support for lexical candidates over time. In Allopenna, Magnuson, and Tanenhaus' (1998) seminal study, participants were asked to follow a spoken instruction (e.g. *Look at the cross. Pick up the candle. Now put it above the square*), while eye movements to four objects on the screen were recorded. The names of some of the objects were phonologically similar to the name of the target object. For example, the target object *candle* was presented with a picture of a competitor that overlapped phonologically with the target word at onset (i.e. *candy*, henceforth an *onset*

competitor) and with a picture of a competitor that rhymed with the target word (i.e. *sandal*, henceforth a *rhyme competitor*). The results showed that in the period before listeners settled unequivocally on the target word, they fixated more on pictures consistent with the target signal (*candy*, *sandal*) than those that were phonologically unrelated to the target signal (*lemon*). Importantly, participants looked more often at onset competitors (*candy*) than at rhyme competitors (*sandal*). These findings suggest that information at the beginning of a word is more important in constraining lexical selection than at the end of the word. In the present study, we examine the modulation of onset and rhyme competition by two types of signal degrading environmental noise that are commonly present in the background: broadband noise and competing speech.

Two studies that examined the influence of non-speech noise on phonological competition used a variant of Allopenna et al.'s (1998) design (Ben-David, Chambers, Daneman, Pichora-Fuller, Reingold, & Schneider, 2010; McQueen & Huettig, 2012). Ben-David et al. (2010) examined age-related differences in spoken word processing and found that both younger and older listeners experienced significantly more competition from an onset competitor in noise than in quiet. Furthermore, older adults experienced more competition from rhyme competitors in a noisy background whereas the younger adults showed only a small negative effect of background noise for rhyme competitor trials. McQueen and Huettig (2012) compared spoken word recognition in quiet with a condition in which the onset phonemes of some of the words in target sentences were replaced with radio-signal noises. The important result of this study was that the listeners' preference for onset competitors was reduced and the preference for rhyme competitors was stronger and occurred earlier in the noise condition than in the quiet condition, suggesting that when onset information is less reliable, listeners seem to adjust their interpretation of the

4

acoustic signal accordingly. Results from both of these studies suggest that the dynamics of spoken word recognition, particularly the relative strength of onset and rhyme competition, are modulated by the presence of background noise.

The current study examined the presence of broadband noise (Experiment 1) and the presence of competing speech signals (Experiment 2) in the background on spoken word recognition. In Experiment 1, we used a primarily energetic masker (broadband noise) which reduces the intelligibility of the target signal due to spectro-temporal acoustic overlap between the target and the background signal. This type of noise does not contain linguistic information. In Experiment 2, the background signal contained intelligible and meaningful content (i.e. speech from a competing talker). This type of background signal can, in addition to providing energetic masking, result in informational masking. In this case, the target and the background signal may both be audible, and therefore difficult to separate with the background signal providing an additional, auditory source of lexical competition during recognition of the target word.

Experiment 1: Broadband noise

Participants

Twenty-five monolingual American-English listeners (8 males, 17 females, age range of 18;9 - 29;8 years) with normal hearing and with normal or corrected-to-normal vision were tested.

Materials

Stimuli were adapted from Allopenna et al. (1998). Sixteen disyllabic nouns referring to picturable objects were selected as targets (see the Appendix for the full set of experimental items). Visual displays consisted of line-drawings of the target object and of two phonologically

related competitor objects. The onset competitor matched with the onset of the target (e.g. *candy* for the target *candle*) and the rhyme competitor matched with the offset of the target (e.g. *sandal* for the target *candle*). The target and competitor overlapped minimally with one syllable. On a given trial, the phonologically related objects were displayed with a distractor object which was phonologically dissimilar to the target and competitors (e.g. *lemon*). Each visual display was presented along with an auditory stimulus ("candle") spoken by an adult female native-English speaker (16 bits, 22050 Hz). The targets (set to play out at 65 dB SPL) were mixed with broadband noise (set to play out at 67 dB SPL) at a SNR level of -2 dB. Targets and noise were presented simultaneously and diotically. An additional 16 and 8 quadruplets were selected for filler and practice trials, respectively.

Procedure, design, and analysis

Prior to the eye-tracking experiment, participants were shown pictures of the objects they were to see in the eye-tracking experiment and asked in a two-alternative forced choice task which of two printed words represented the picture. Results showed that participants made no errors, indicating that our pictures were highly identifiable.

Participants' eye-movements were monitored at a sampling rate of 1 kHz with an SR Research EyeLink1000 eye-tracker. The presentation of the auditory and visual stimuli was controlled with Experiment builder. The auditory stimuli were presented over headphones. After a calibration procedure, participants received written instructions on the screen. They were asked to click on the picture in the visual display representing the target word they heard. We presented two conditions which were blocked: the quiet condition always preceded the noise condition. Each block consisted of a total of 8 experimental and 8 filler trials. The trials were randomized.

Before each block a practice session of 4 trials was presented. The location of the pictures was randomized over the four quadrants of the screen. On each trial, the four pictures were first displayed. After 1000 ms, the auditory stimulus was presented. When participants clicked with the mouse on a picture, they initiated the next trial. The experimental session lasted about 15 minutes.

A statistical analysis of the error pattern, the response times (RTs) and the eye movements was carried out with linear mixed effects models (Baayen, Davidson & Bates, 2008). For the click responses, the percentage of correct identifications was calculated. The RTs on the correct detections were measured from target word offset. Following Allopenna et al. (1998) fixation proportions in non-overlapping 100 ms time bins from 200 to 800 ms were analyzed in order to provide a fine-grained description of how competition emerged over time. Fixations were transformed into empirical logits (Barr, 2008). From these data, we created three measures: 1) mean of looks to the onset competitor vs. the distractor (onset competition; onset-distractor); 2) means of looks to the rhyme competitor vs. the distractor (rhyme competition; rhyme-distractor); and 3) means of looks to the onset vs. the rhyme competitor (specific competition, onset-rhyme). In the model Condition (quiet vs. noise) was entered as fixed effect and participants and items as random effects. Condition was coded as a numeric contrast, that is, the quiet condition as -0.5 and the noise condition as +0.5. A negative regression weight (beta) implies more fixations in the quiet than in the noise condition. In the Results section, we report the beta range (including respective p-values) across the significant time bins. The p-values are estimated by using Markov chain Monte Carlo simulations. Baayen et al. (2008) has suggested this as an adequate way to deal with smaller datasets.

Results and Discussion

The accuracy analysis showed that listeners performed significantly better in quiet (99%) than in noise (91%; $\beta_{condition}$=-1.10; p<0.003). The response time data revealed that listeners needed significantly more time to respond in noise (M=997 ms, SD=448 ms) than in quiet (M=793 ms, SD=316 ms; $\beta_{condition}$=207.9; p<0.0002). Figure 1 shows the proportion of fixations over time from 0 to 1200 ms after target word onset for the correct trials in the A) Quiet condition and B) Noise condition. The analysis showed onset competition (relative to the distractor) from 200 to 700 ms ($\beta_{Intercept}$ ranged from 0.38 (p_{MCMC}<0.04) to 0.51 (p_{MCMC}<0.02)). This pattern was dependent on Condition from 400 to 700 ms ($\beta_{Condition}$ ranged from 0.59 (p_{MCMC}<0.04) to 0.99 (p_{MCMC}<0.002)), indicating that onset competitors attracted more looks (relative to the distractor) in the noise than in the quiet condition. An additional analysis in each condition separately showed that there was a preference for the onset competitor only in the noise condition. The analysis also showed rhyme competition (relative to distractor) from 400 to 800 ms ($\beta_{Intercept}$ ranged from 0.36 (p_{MCMC}<0.04) to 0.48 (p_{MCMC}<0.01)). This pattern was dependent on Condition from 700 to 800 ms ($\beta_{Condition}$=0.81; p_{MCMC}<0.005). An additional analysis in each condition separately indicated that there was only a rhyme competitor preference (relative to the distractor) in the noise condition. Finally, the specific competition analysis (onset-rhyme) showed that looks to onset competitors never differed from rhyme competitors.

In sum, the eye-tracking data of Experiment 1 revealed onset relative to distractor competition in both the quiet and the noise condition in an early time window, and this type of competition persisted for a longer time window in the noise condition than in the quiet condition. Late in time there was persistent competition from both onset and rhyme competitors in the broadband noise condition, indicating that overall competition is enhanced throughout the period of lexical

selection when target uncertainty increases due to the presence of background noise. Consistent with the findings from Huettig and McQueen (2012) who reported an increase in rhyme competition effects under conditions of uncertainty; these findings suggest that the dynamics of spoken word recognition are modulated by broadband noise. When target information is less reliable, listeners interpret the acoustic signal with more flexibility entertaining competing lexical items for a fairly long period of time as the word unfolds rather than quickly eliminating all lexical competitors. The present results are inconsistent with Ben-David et al.'s (2010) results, who found only a small negative effect of background noise on rhyme competition for younger listeners. However the younger listeners in that study experienced significantly more competition from the onset competitor in noise than in quiet. Thus, while they differ in the details, this study and the present experiment both show greater competition overall in noisy than in quiet listening conditions. Possible reasons for the discrepancies are their use of sentence rather than word stimuli and of displays with one competitor per trial instead of displays with both onset and rhyme competitors within each trial.

In Experiment 2, we tested how words are recognized when there is competition from the mental lexicon as triggered by the target signal in addition to competition from the auditory modality as triggered by background speech presented concurrently with the target speech. We chose to present one word (produced by a target talker) mixed with another word (produced by a background talker) to create as much of a controlled manipulation as possible in a speech-in-speech situation. It is thus a first step towards obtaining ecological validity, although in real-world speech communication the likelihood of hearing two talkers synchronized is very small. Furthermore, this arrangement of stimuli is likely to stress the processes of stream segregation quite severely thereby maximizing the chances of observing an effect. We presented an onset

competitor concomitantly in the background in an onset background speech condition (e.g. *candle-candy*) and a rhyme competitor concomitantly in the background in a rhyme background speech condition (e.g. *candle-sandal*).

Experiment 2: Competing speech

Participants

Twenty-six monolingual American-English listeners (7 males, 19 females, age range of 18;6 - 24;3 years) with normal hearing and normal or corrected-to-normal vision participated.

Materials

The same materials were used as in Experiment 1, except that in Experiment 2 we removed the quiet condition and we replaced the broadband noise with naturally produced words. The visual display was presented along with a target word (e.g. *candle*) mixed with a background word as described below. The background word was either an onset competitor (e.g. *candy*) or a rhyme competitor (e.g. *sandal*). We recorded an additional voice of a female, native American-English talker who was different from the target talker. She was designated as the background talker. We manipulated the duration of the words to be of equal length so that participants could not use durational differences as a segregation cue. Words were compressed or lengthened in PRAAT©. The duration manipulations were minimal and therefore produced little or no distortion. The targets were then mixed with the background words. The level of the target words (set to play out at 65 dB SPL) was fixed at a level that was 2 dB higher than that of the background words (set to play out at 63 dB SPL). This difference in SNR helped listeners to segregate the target and the background speech streams.

Procedure, design, and analysis

Participants' task was to attend to the target talker and to ignore the background talker. 8 practice trials familiarized the participants with the task, as well as with the voices of the target and background talkers. Participants were then presented with a total of 32 experimental and filler trials. Onset and rhyme noise trials were randomly mixed. Two different item lists were created. Both lists contained half of the targets in onset background speech and half of the targets in rhyme background speech. Each participant received one list. The trials in each list were randomized. The experimental session lasted about 15 minutes. In the model Background (onset vs. rhyme) was entered as fixed effect and participants and items as random effects. Background was coded as a numeric contrast: the onset condition as -0.5 and the rhyme condition as +0.5. A negative regression weight (beta) implies more fixations in the onset than in the rhyme condition.

Results and Discussion

The accuracy analysis showed that listeners performed better in the onset condition (82%) than the rhyme condition (64%; β=-0.91; p<0.0001). Moreover, listeners clicked 17% of the time on onset competitors in the onset condition. In contrast, listeners clicked 34% of the time on rhyme competitors in the rhyme condition. The response time data showed that listeners needed approximately the same amount of time to click on targets in the onset (M=1114, SD=439) as in the rhyme condition (M=1185, SD=471; p>0.05).

Figure 1 shows the proportion of fixations over time from 0 to 1200 ms after target word onset for the correct trials in the C) Onset condition and D) Rhyme condition. The analysis revealed onset competition (relative to the distractor) between 300 to 800 ms ($\beta_{Intercept}$ ranged

from 0.44 (p_{MCMC}<0.04) to 0.56 (p_{MCMC}<0.03)). This pattern was never influenced by Background, indicating that onset competition was equally strong in both conditions. The analysis also showed late rhyme competition (relative to the distractor) from 500 to 800 ms ($\beta_{Intercept}$ ranged from 0.61 (p_{MCMC}<0.01) to 0.79 (p_{MCMC}<0.0003)). Importantly, this pattern was dependent on Background in the same time bins ($\beta_{Background}$ ranged from 0.83 (p_{MCMC}<0.02) to 1.51 (p_{MCMC}<0.0003)). The positive regression weights indicate that rhyme competitors attract significantly more attention in the rhyme than in the onset condition. An additional analysis in each condition separately showed that there was a preference for the rhyme competitor only in the rhyme condition from 500 to 800 ms. Furthermore, an analysis of specific competition (direct comparison of onset versus rhyme fixations) was dependent on Background from 500 to 800 ms ($\beta_{Background}$ ranged from -0.71 (p_{MCMC}<0.05) to -1.88 (p_{MCMC}<0.0003)). It also reached significance in the 200 to 300 ms time window (p_{MCMC}<0.04). The analysis on each condition separately showed that this effect was significant for the rhyme condition.

In sum, these data show that the dynamics of the spoken word recognition system interact with background speech because lexical competition patterns change as a function of the content of background speech. Listeners pay extra attention to the competitor that matches the background speech in the visual display. For example, hearing *candy* in the background increases looks to the picture of a *candy*, and hearing *sandal* in the background increases looks to a picture of a *sandal*. This indicates that competition from a simultaneously heard rhyme competitor can change the previously demonstrated onset competition advantage (e.g. Allopenna et al., 1998). The competitor set is thus not only determined by the target input, but also by what occurs in the auditory background. Put another way, the specific lexical content of to-be-ignored

background speech influences patterns of lexical competition throughout the time course of spoken word recognition.

General Discussion

This study examined the temporal dynamics of spoken word recognition in background noise and background speech. In Experiment 1, we compared recognition of target words in quiet and in broadband noise. Result showed that under adverse listening conditions competition was strong and persisted until late in time. Importantly, competition from the rhyme competitor (e.g. looking at *sandal* when hearing *candle*) increased compared to the quiet condition indicating that the presence of broadband noise changed the dynamics of spoken word recognition. Listeners were less certain as they heard speech in noise, thereby increasing looks to phonological competitors. This uncertainty lasted until a late moment in time, revealing an overall and persistent processing cost in noise even for the correctly identified trials.

Experiment 2 compared the influence on the competition process of background words that matched the target word (e.g. *candle*) either in terms of their onsets (e.g. *candy*) or their rhymes (e.g. *sandal*). Results showed that the lexical competition process changed as a function of the background speech. For correctly identified trials, when listeners heard *candy* (onset match) in the background, their eye gaze shifted to the picture of a *candy*, but when listeners heard *sandal* (rhyme match) in the background, their eye gaze shifted to the picture of a *sandal*. In the latter case, the strong preference to look at an onset competitor, as usually found in this type of work (e.g. Allopenna et al., 1998), was reduced. This finding reveals that the spoken word recognition system adjusts to the auditory input it receives. This adjustment is not only based on the target input, but also on the input from the background. Similarly, Brouwer, Mitterer and Huettig

(2012) have recently shown that the competition process may also change depending on whether listeners hear the same target word produced canonically (e.g. *computer*) or in a reduced way (e.g. *puter*). In that study, they found that listeners had a preference to look at onset competitors when all targets were fully pronounced, but not when these fully pronounced targets were intermixed with reduced targets. This suggests that listeners penalize acoustic mismatches less strongly when the listening context as a whole is non-optimal by including reduced speech.

It is possible that our results do not show differences between onset and rhyme competition per se, but that the critical factor is the amount of segmental overlap between the target and the background signal. We therefore analyzed to what extent onset and rhyme competitors differ phonetically from their target words. We calculated how many segments were shared between the competitors and their targets. The number of matching segments was then divided by the total number of segments of the onset competitor. Similar comparisons were made between the rhyme competitors and their targets. A paired t-test showed a significant difference in segmental overlap between the values for the onset and the rhyme competitors (averages of 0.55 and 0.68 for onset and rhyme, respectively, $t(15)=-2.411$, $p<0.05$), indicating that the rhyme competitors were more phonetically similar than the onset competitors. This finding is consistent with the idea that the amount of segmental overlap between the target and the background signal might drive the specific competition patterns that we observed. The accuracy data are in line with this pattern. It was harder to ignore rhyme competitors than onset competitors in the background signal. Nevertheless, regardless of whether the observed differences are due to the type (rhyme or onset) or amount (number of segments) of target-to-background overlap, the results from Experiment 2 indicate that lexical competition in the visual world paradigm is highly sensitive to

the presence of background speech throughout the observation window (all the way out to 800 msec post stimulus onset).

A critical aspect of spoken word recognition in noisy listening conditions is that, in addition to the processes of lexical selection, speech-from-noise segregation is also required. Based on the present data, we suggest that a discrete-serial model of stream segregation and lexical selection could be rejected because we found dramatically different (not just delayed) patterns of lexical competition as seen in the gaze patterns in the visual world paradigm across all noise and background speech conditions (see Rapp & Goldrick, 2000 for a discussion on discreteness and interactivity). There is apparently no point in time when stream segregation ends and the typical pattern of lexical competition begins. Instead, our data show gaze patterns that are highly sensitive to the content of the auditory background throughout the observation window, and are therefore more compatible with non-discrete, interactive accounts of stream segregation and lexical selection. That is, lexical selection and stream segregation appear to be temporally integrated rather than temporally segregated.

It is important to note, however, that a strictly sequential account of the present data could possibly be upheld by supposing that, after stream segregation is completed, the signal available for the processes of lexical selection to act upon is degraded in adverse relative to quiet listening conditions. Under this view, lexical items presented in adverse conditions are never as robustly distinguished from their competitors as lexical items heard under ideal listening conditions thereby accounting for the observed lower rates of correct identification and greater competition throughout the observation time window even for correctly identified trials. Moreover, it is possible that listeners looked longer at the competitors because they were entertaining the

hypothesis that the competitors were the targets. The eye movement pattern would in this case present a failure of selective attention and/or stream segregation.

Finally, an important issue that remains to be investigated is the extent to which our perfectly synchronized target and background signal in Experiment 2 are ecologically valid. Moreover, in the present study we used background words that were carefully selected to match either the onset or rhyme of the target word. This carefully synchronized and highly controlled target-to-background relationship emphasized the impact of the background signal on processing of the target signal, however this is a situation that is rarely (if ever) encountered in the real-world. Thus, for a comprehensive understanding of the interplay between stream segregation and lexical competition, a next step would be to see if similar patterns could be observed under more realistic listening conditions. Note, however, that in Experiment 1 where the background signal was broadband noise (representing a fairly common real-world, adverse listening environment) we also observed stronger and more persistent lexical competition than in the quiet condition, indicating that regardless of the content of the background noise, its impact on spoken word recognition is evident throughout the period of time leading up to final word selection. A challenge for future work is to more finely delineate the influence of different types of background noise where the extent of energetic and non-energetic (i.e. informational or linguistic) overlap between the target and background signals varies.

To conclude, this work contributes to the discussion of how listeners cope with variability in speech as caused by extrinsic noise and background speech. Overall the results suggest that stream segregation and lexical competition may be temporally integrated because competition changes as a function of the presence and the specific (lexical) content of the background signal.

Acknowledgments

References

Allopenna, P.D., Magnuson, J.S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419-439.

Baayen, R.H., Davidson, D.J., & Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390-412.

Barr, D. J. (2008). Analyzing 'visual world' eye tracking data using multilevel logistic regression. *Journal of Memory and Language, 59*, 457-474.

Ben-David, B.M., Chambers, C.G., Daneman, M., Pichora-Fuller, K.M., Reingold, E.M., & Schneider, B. (2010). Effects of aging and noise on real-time spoken word recognition: Evidence from eye movements. *Journal of Speech, Language and Hearing Research, 54*, 243-262.

Brouwer, S., Mitterer, H., & Huettig, F. (2012). Speech reductions change the dynamics of competition during spoken word recognition. *Language and Cognitive Processes, 27*(4), 539-571

McQueen, J.M. (2007). Eight questions about spoken-word recognition. in *The Oxford handbook of psycholinguistics,* edited by M. G. Gaskell, (Oxford: Oxford University Press), pp. 37-53.

McQueen, J.M., & Huettig, F. (2012). Changing only the probability that words will be distorted changes how they are recognized. *Journal of Acoustical Society of America, 131*(1), 509-517.

Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review, 107*(3), 460-499.

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., & Sedivy, J.C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632-1634.

Figure 1: Proportion of fixations over time from 0 ms till 1200 ms after target word onset to targets, background and target competitors, and distractors in Experiment 1 (A: Quiet condition, B: Noise condition) and Experiment 2 (C: Onset condition, D: Rhyme condition).
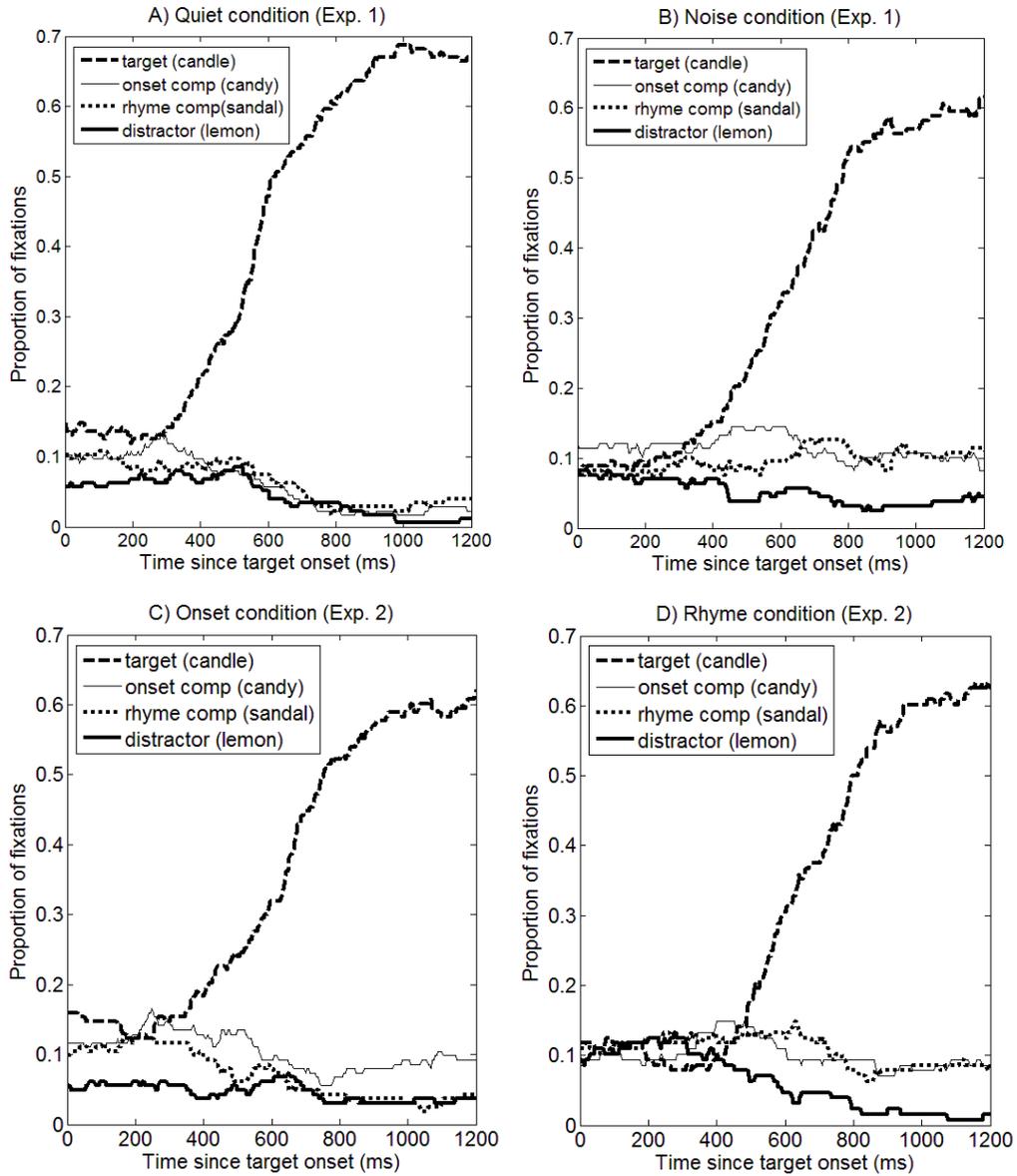
Appendix: List of experimental items.

	Target	Onset comp.	Rhyme comp.	Distractor
1	beaker	beetle	speaker	zebra
2	button	butter	curtain	tiger
3	candle	candy	sandal	lemon
4	carrot	carriage	parrot	helmet
5	casket	castle	basket	airplane
6	dollar	dolphin	collar	panda
7	kitten	kitchen	mitten	bottle
8	letter	lettuce	sweater	trumpet
9	peanut	peacock	walnut	dresser
10	pedal	pepper	medal	statue
11	pencil	penguin	stencil	hammer
12	pillow	pillar	aloe	balloon
13	racket	rattle	jacket	mirror
14	rooster	ruler	hamster	bucket
15	tower	towel	mower	apple
16	turkey	turtle	monkey	ankle