The Effects of Onset and Offset Masking on the Time Course of Non-Native Spoken-Word Recognition in Noise

Anonymous CogSci submission

Abstract

Using the visual-word paradigm, the present study investigated the effects of word onset and offset masking on the time course of non-native spoken-word recognition in the presence of background noise. In two experiments, Dutch non-native listeners heard English target words, preceded by carrier sentences that were noise-free (Experiment 1) or contained intermittent noise (Experiment 2). Target words were either onset- or offset-masked or not masked at all. Results showed that onset masking delayed target word recognition more than offset masking did, suggesting that – similar to natives – non-native listeners strongly rely on word onset information during word recognition in noise.

Keywords: spoken-word recognition; non-native listeners; background noise; visual world paradigm

Introduction

Compared to speech processing under ideal circumstances, recognizing words in the presence of background noise is substantially more difficult (Mattys et al., 2012). The obvious problem is that the certainty with which words can be recognized decreases because noise masks relevant phonemic elements, often at non-predictable moments in time. A compelling body of research has shown that speech processing in noise is especially hard in a non-native language (Garcia Lecumberri et al., 2010).

Previous experimental studies on native listeners have often used the visual world paradigm (Huettig et al., 2011) to investigate how noise affects the real-time dynamics of spoken-word recognition. For example, noise-induced ambiguity in the speech signal has been suggested to result in enhanced competition between the target and similar sounding target word candidates (e.g., Ben-David et al., 2011; Brouwer & Bradlow, 2016). Moreover, McQueen and Huettig (2012) showed that intermittent noise that masked speech preceding a target word led listeners to change the relative weighting they gave to word-initial versus word-final information when recognizing the target, which itself was not masked.

In addition to affecting the competition dynamics, noise has dramatic effects on the time course of spoken-word recognition. McMurray and colleagues (2017) measured how quickly participants made eye movements to objects that were mentioned in the speech signal. They found that targets were looked at more than 200 ms later when speech was degraded relative to when speech was clear. McMurray and colleagues reasoned that listeners might adopt a 'wait-and-see' strategy in the face of signal degradation resulting in delayed lexical access. Note that such behavior is in stark contrast with standard models of spoken-word recognition, which assume that listeners engage in lexical access (and experience lexical competition) as early as possible.

Little is known about how the presence of background noise affects the dynamics of *non-native* spoken-word recognition. A recent study took an important step towards filling that gap in the literature. XXX (YYYY) investigated the relative importance of word-onset and word-offset information when listening in noise. They tested Dutch non-native listeners of English on a word transcription task, where English target words were presented to the participants either in the clear or masked by speech-shaped background noise at different signal-to-noise ratios (SNRs). Critically, when presented in noise, either the word onset or the word offset was masked. The authors found – as to be expected – that overall transcription accuracy dropped as SNRs decreased (i.e., listening became harder). In line with previous studies on native listeners, analyses of participants' misperceptions suggested that noise increased the competitor space (Ben-David et al., 2011; Brouwer & Bradlow, 2016). Moreover, the authors observed that onset masking had a more detrimental effect than offset masking, demonstrating a strong reliance on word onset information when recognizing spoken words.

Arguably, as the critical measure in XXX (YYYY) was collected after spoken-word offset, additional reasoning and decision processes may have contributed to participants' transcriptions. Thus, it is not known how onset and offset masking affect *online* spoken-word recognition in non-native listeners. In the present study, we followed-up on the study by XXX (YYYY) using an online task. Specifically, we used the visual world eye-tracking paradigm to study the effects of word onset and offset masking on the time course of non-native spoken-word recognition. To that end, we augmented

the target words used by XXX (YYYY) with semantically-neutral carrier sentences and presented them to Dutch non-native listeners of English. Target words were either presented in the clear, or target word onsets and offsets, respectively, were masked with speech-shaped background noise at different SNRs. While listening to the sentences, participants looked at displays featuring four objects. On target-present trials, a picture of the target word was shown, along with three unrelated distractors. On target-absent trials, the picture of the target word was absent and participants saw two pictures, whose word names overlapped phonologically with the unfolding target, and two unrelated distractors.

Our analyses focused on target-present trials to address how the time course of non-native spoken-word recognition is affected by onset and offset masking and different SNRs. We consider the moment in time when participants displayed a significant bias for the target objects (i.e., more looks to the target than to the unrelated distractors) to reflect that they exhibit phonological mapping of the incoming speech signal onto the names retrieved from previewing the pictures (i.e., lexical access). In Experiment 1, the carrier sentences were free of any noise. The carrier sentences in Experiment 2 contained intermittent noise, that is, noise masking random parts of the sentence, similar to a badly tuned AM radio (McQueen & Huettig, 2012). We used this manipulation to investigate whether listeners adjust their reliance on onset or offset information when there is a high probability of the target word being masked.

If the presence of noise leads non-native listeners to adopt a 'wait-and-see' strategy (as suggested for native listeners, McMurray et al., 2017), we should in general observe no difference in the detrimental effects of offset masking as compared to onset masking: if participants waited for 200 ms (or more) after word onset before engaging in lexical access, the relative position of less reliable information should become less important. That is, listeners' reliance on word onsets would be weakened. In contrast, if listeners immediately engage in mapping the incoming speech onto the retrieved picture names, onset masking should have a more detrimental effect than offset masking.

For Experiment 2, if the high probability of the target being masked (induced by the presence of intermittent noise distributed over the carrier sentences) leads participants to rely more strongly on word offset information and less strongly on word onset information than when listening in the clear (as suggested for native speakers, McQueen & Huettig, 2012), the presence of intermittent noise should result in a smaller difference between the onset and offset masking conditions compared to Experiment 1.

Importantly, for both experiments our focus was not on comparing the absolute numbers (points in time) of target word biases to those reported in other studies. Instead, we were interested in the (change of the) relative importance of word onset and offset information in the situations tested in the present experiments. Our clear-speech conditions constituted experiment-internal baselines against which the various noise conditions were compared.

Experiment 1

Participants

Twenty native speakers of Dutch (14 females; M age = 22.6; years, SD = 3.1, range = 18-33), students at XXX University (YY), participated in Experiment 1. They were given a voucher for their participation. None of the participants reported a history of speech and/or hearing disorders or a history of neurological problems. All participants signed a consent form prior to the experiment.

All participants were proficient users of English as assessed using LexTALE (Lexical Test for Advance Learners of English; Lemhöfer & Broersma, 2012), which was administered after the main experiment. LexTALE is an easy-to-use, swift vocabulary knowledge test, based on a visual lexical-decision task, consisting of 60 trials, 40 English words and 20 non-words presented in random order. The test takes about 3 to 4 minutes to complete. Participants' mean LexTale score in Experiment 1 was 76 (*SD* = 10.28, range = 61 - 93), which corresponds to an upper intermediate (B2) proficiency level of English (Lemhöfer & Broersma, 2012).

Method

Materials We adjusted the materials from XXX (YYYY) for use in a visual world eye-tracking experiment. We selected 80 of their targets; all were concrete English nouns. The frequency of these words, operationalized as Zipfian frequency (van den Heuven et al., 2014), was high (M = 4.37, one word was not listed). Target words were on average 3.7 phonemes long.

We took the recordings used by XXX (YYYY). Those had been produced by a male native speaker of Southern British English and had been recorded in a sound-attenuated booth at 44.1 kHz. Using Praat (Boersma, 2011), stationary speech-shaped background noise (SSBN) had been added to the files. To do so, XXX (YYYY) down-sampled each target word recording to 16 kHz to make them compatible with a sound file containing SSBN. A random stretch of SSBN was automatically selected from the file and was placed on the target word, applying a Hamming window, with a 10 ms fade in/out. The average (i.e., root-mean-square) intensity was set to 60 dB SPL. SSBN was added at two different SNRs: -6 dB and -12 dB. In XXX (YYYY), these SNRs had neither yielded ceiling nor floor effects in terms of transcription accuracy. Moreover, there was a substantial accuracy difference between both conditions. Importantly, the authors placed SSBN either on the onset or on the offset of the target word. This was done by using a semi-automatic method, where first the boundaries for onsets (M = 2.5 phonemes) and offsets (M = 2.7 phonemes), respectively, in each target word were manually set at positive-going zero-crossings. Then, noise was added to the stretch spanning either onset or offset, making these harder to understand. In total, there were five versions of each target word (i.e., constituting the five conditions) in the present experiments: four noise versions (onset and offset masked at SNR -6 and -12 dB) and a clean (noise-free) version.

Each target word was combined with one of eight different carrier sentences and appeared as the sentence-final word. The carrier sentences were semantically neutral and not predictive of the target word (e.g., "As her password she chose *bed*"). We asked the same person who had previously produced the target words to produce the carrier sentences (without the target words at the end) and made recordings using the same parameters as before. The mean duration of the carrier sentences was 1825 ms. The mean duration of the target words on target-present trials was 560 ms; that of the target words on target-absent trials was 518 ms.

The 80 items were evenly divided into target-present and target-absent trials. All objects were black-on-white drawings selected from the Snodgrass and Vanderwart (1980) database.

Procedure Participants were tested individually in a sound-attenuated booth. Auditory stimuli were presented binaurally over closed headphones. Eye movements were recorded using an EyeLink 1000 eye-tracker sampling at 1000 Hz. The visual stimuli were displayed in an array on the computer screen that spanned 1024 x 768 pixels.

Following presentation of the instructions, the eye-tracker was calibrated. Participants performed a look-and-listen task (Huettig et al., 2011), which meant that they could look wherever they wanted while not taking their eyes off the screen. This task is assumed to reduce the engagement of additional reasoning and decision processes. The trial structure was as follows: A fixation dot appeared in the center of the screen for 250 ms, followed by the presentation of the four objects. After one second, the playback of the carrier sentence started. Immediately after the carrier sentence had ended, the playback of the target word started. The four objects remained in view until the end of the trial. The inter-trial interval was one second.

The 40 target-present and 40 target-absent items were rotated across the five listening conditions (while the visual objects stayed the same) such that one participant heard a given target only once: either onset or offset masked, at an SNR of -6 or -12 dB, or without noise. This resulted in five experimental lists with 80 trials each. Trial presentation on the lists was blocked by listening condition, with the clean condition always coming last (to mitigate fatigue effects from listening in noise). Moreover, the two onset-masking and the two offset-masking conditions always followed each other such that SNR -6 dB always preceded SNR -12 dB (to allow participants to get used to the noise). We counterbalanced the order of masking conditions such that half of the participants heard onset-masked targets first and the other half heard offset-masked targets first.

Participants were assigned to one list. The five blocks on each list each contained 16 trials, half of which were target-present and half were target-absent, presented in random order. Participants were presented with all 80 trials on the list; they could take short breaks between blocks. Breaks were followed by a drift check to ensure that the tracker was still well calibrated.

Data Analysis Target-present trials for the five listening conditions were processed for a period starting at target word onset and ending 1500 ms thereafter. Track loss was excluded. The processed data were analyzed using a logistic additive mixed model using function *bam* from R package mgcv (Wood, 2017). The dependent variable was the fixation object, which was either target (1) or distractor (0). A fixed-effect predictor was included for Condition (treatment-coded with 'clean' as the reference). The temporal trajectory of the target fixations over time was modeled using a thin-plate regression spline, which was afforded at most 20 basis functions, by each condition. Random effects were added for Condition by participants and by items, as were random smooths of the aforementioned thin-plate regression spline by subjects and by items.

The average of the three distractors was included as an offset term, such that the estimated fixation proportions were relative to a chance level corresponding to this average, rather than to a fixed proportion of .5. An autoregressive error process of order 1 was included with $\rho = .9$ to account for the fact that adjacent samples in time were not independent. To mitigate convergence failures, which were due to complete separation causing indefiniteness in the likelihood, we added a small smoothing constant of .01 to the target and the non-target looks, following Donnely and Verkuilen (2017).

Significance of the fixation differences between the target and the average of the distractors over time (i.e., target bias) was established by computing Bayesian 95% credible intervals, following Wood (2007, pp. 293-294). Target preferences were considered significant where the CI excluded zero on the log-odds scale, which corresponds to the average of the distractors on the proportion scale.

Results

Participants' fixation proportions are plotted in Figure 1; the fitted trajectories are shown in Figure 2. In all five conditions, participants showed an extended fixation bias for the target objects over the unrelated distractors, which suggests that they recognized the spoken non-native target words. To examine the time course of target word recognition across the five listening conditions, we focus on the moments in time – starting from target word onset – where the difference between looks to the target and the distractor objects reached statistical significance. As it takes minimally 200 ms (Saslow, 1967) to program and launch a saccadic eye movement, we can consider gaze 200 ms after target word onset to reflect processing of the target word. Recall that the mean target word duration (on target-present trials) was 560 ms. Thus, target biases occurring before 760 ms are assumed to reflect effects during the target words' unfolding.

In the clean condition, participants fixated the target objects (more than the unrelated distractors) shortly after they were mentioned, reaching significance at 464 ms after target word onset. The time course was similar when offset information was less reliable due to the presence of SSBN at

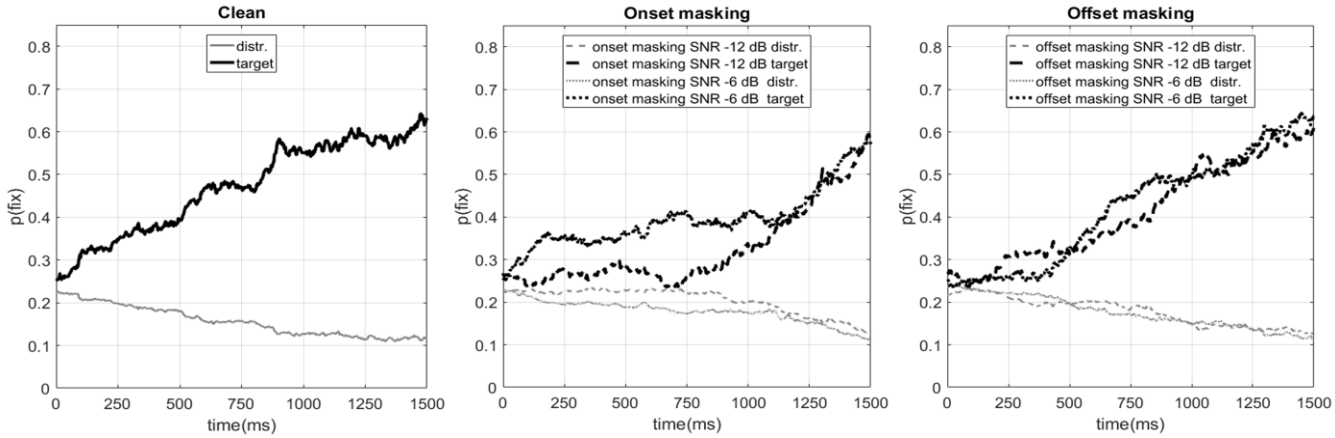

Figure 1: Fixation proportions of clean, onset- and offset-masking (at SNR -6 and -12 dB) conditions in Experiment 1.

SNR -6 dB: The target fixation bias in the offset masking condition reached statistical significance at 483 ms after target word onset. In the SNR -12 dB offset-masking condition, the target bias occurred at 626 ms after target word onset. When onset information was less reliable, the same target objects showed a later fixation bias: 534 ms in the SNR -6 dB and 895 ms in the SNR -12 dB condition, respectively.

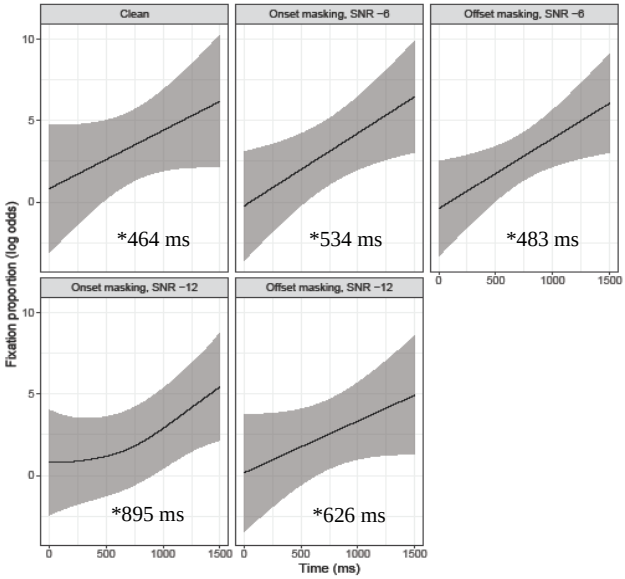

Figure 2: Fitted temporal trajectories for Experiment 1, controlling for random effects, relative to the average of the distractors. The gray bands show the 95% CI. The number in each panel reflects the point in time when the target bias became significant.

Discussion

Our analyses showed that participants fixated the target objects more than the unrelated distractors in all listening conditions, demonstrating that they recognized the target words. As to be expected, the target bias occurred earliest in the clean condition. Note that compared to native listeners, who typically show target biases 200 ms after word onset (cf. McMurray et al., 2017), a somewhat later bias is expected when engaging in non-native lexical access (cf. Garcia Lecumberri et al., 2011). In the offset masking condition, at SNR -6 dB, we observed a similar time course, which differed from the clean condition by only a few milliseconds. Masking word onsets was much more detrimental than masking word offsets. In both SNR conditions, the target bias occurred later than in the corresponding offset masking conditions. Moreover, the difference between -6 and -12 dB SNR conditions was more pronounced in the onset masking conditions—differing by 360 ms (vs. a difference of 143 ms in the offset masking conditions).

The similar time courses of the clean and offset SNR -6 dB conditions and the more detrimental effects of onset masking suggest a stronger reliance on word onset than offset information when listening in noise. Non-native listeners engaged in mapping the incoming speech signal onto the retrieved phonological codes as early as possible. These results do not provide evidence for a 'wait-and-see' strategy.

In sum, using an eye-tracking paradigm assumed to tap into *online* spoken-word recognition, the data from Experiment 1 are in line with the results described in XXX (YYYY), who used an *offline* transcription task. We found that masking word onset information led to considerably later target biases than masking offset information suggesting a stronger reliance on onset than offset information.

In Experiment 2, we used a design that has previously been shown to change listeners' relative reliance on word-offset versus word-onset information in native listeners (McQueen & Huettig, 2012). The crucial question was whether – compared to Experiment 1 – we would observe more similar effects of onset and offset masking.

Experiment 2

Participants

Twenty native speakers of Dutch (16 females; M age = 22.3 years, SD = 2.5, range = 18-28), students at XXX University (YY), participated in Experiment 2. They were given a voucher for their participation. None of the participants reported a history of speech, hearing disorders and/or

neurological problems. All participants signed a consent form prior to the experiment. As for Experiment 1, participants, completed the LexTale task (Lemhöfer & Broersma, 2012). Their mean LexTale score was 80.5 (*SD* = 13.2), which corresponds to proficient or C1 proficiency-level of English (Lemhöfer & Broersma, 2012).

Method

The method was the same as in Experiment 1 except that the carrier sentences contained intermittent noise (following McQueen & Huettig, 2012). The intermittent noise consisted of short stretches of noise that were taken from the SSBN file used for the noise added to the target words. These stretches had the same duration as the noise added to the target words and had the same SNR. The stretches were placed at random positions in the carrier sentence using a custom-made Praat script. They always occurred at positive-going zero crossings. Each of the eight carrier sentences received between two and five stretches of noise.

Results

Participants' fixation proportions are plotted in Figure 3; the fitted trajectories are shown in Figure 4. In all five conditions, participants showed a fixation bias for the target objects over the unrelated distractors, suggesting that they recognized the spoken non-native target words. In the clean condition, target object fixation reached significance at 681 ms after target word onset. The target fixation bias in the onset masking condition reached statistical significance at 848 ms after target word onset in the -6 dB condition and slightly later at 928 ms after word onset in the -12 dB condition. The target fixation bias in the offset masking condition reached statistical significance at 732 ms after target word onset for the -6 dB condition but already at 414 ms after word onset for the -12 dB condition.

Discussion

Compared to Experiment 1, all target biases in Experiment 2 occurred later, which is most likely associated with the presence of intermittent noise in the carrier sentences. There was one exception: The -12 dB offset masking condition showed a bias that occurred even earlier than that of the clear condition in Experiment 1. This finding is unexpected and in contrast with our hypotheses. We inspected the by-trial data of that condition, but could not identify a clear source of this effect. It is especially puzzling that the same offset-masked words yielded quite different results in Experiment 1. We therefore refrain from interpreting this condition.

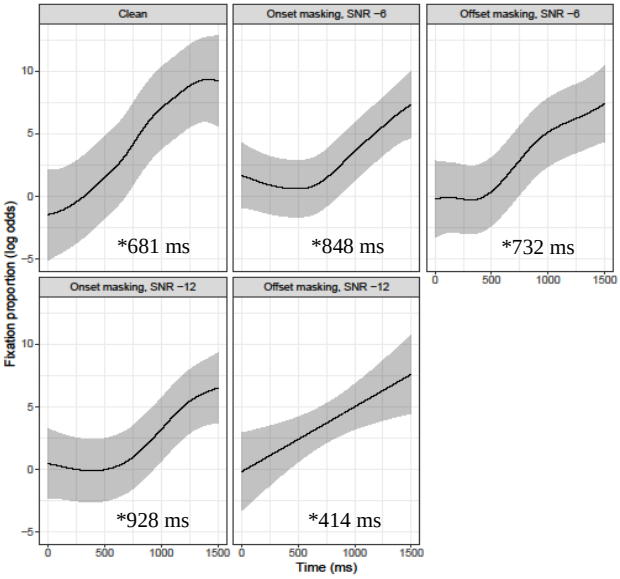

Figure 4: Fitted temporal trajectories for Experiment 2, controlling for random effects, relative to the average of the distractors. The gray bands show the 95% CI. The number in each panel reflects the point in time when the target bias became significant.

With regards to the remaining conditions, we observed that – as before – target bias occurred earliest in the condition without any noise on the target word. In the SNR -6 dB conditions, we observed again that onset masking led to a later target bias than offset masking (by almost 100 ms) and

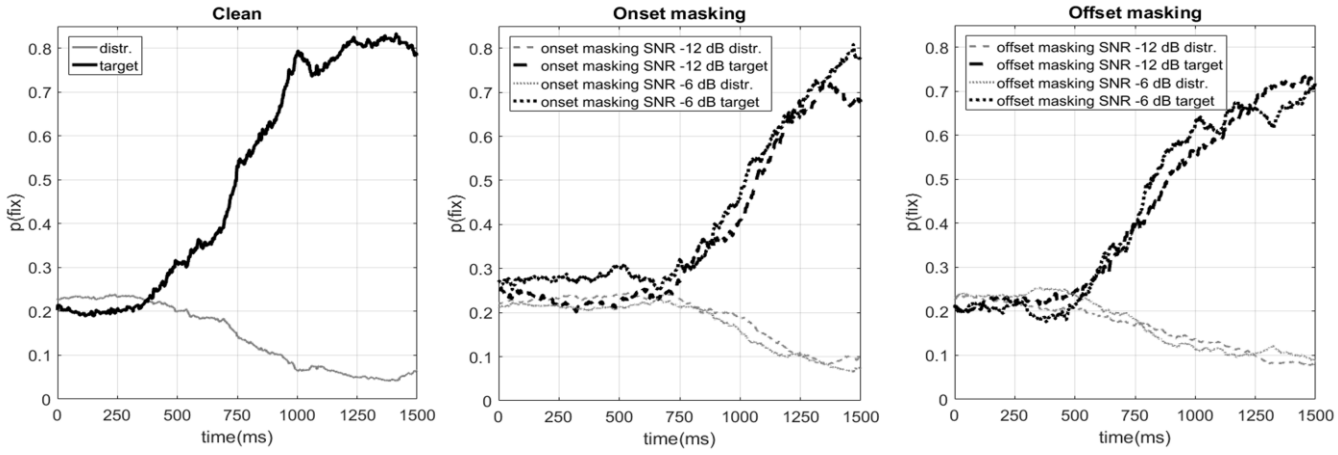

Figure 3: Fixation proportions of clean, onset- and offset-masking (at SNR -6 and -12 dB) conditions in Experiment 2.

that the -6 dB offset masking condition differed from the clean condition by only 51 ms. Note that the biases in the onset masking conditions (at both SNRs) occurred after target word offset and hence after the period that would reflect processing effects during the word's unfolding (i.e., 760 ms after target onset). Compared to Experiment 1, the difference between the two onset-masking conditions was smaller in Experiment 2.

In sum, adding intermittent noise to the carrier sentences might have driven non-native listeners towards a 'wait-and-see' approach, as reflected in the overall later target biases in Experiment 2 compared to Experiment 1. However, it did not affect their tendency to rely more on word onset than offset information as more detrimental effects were observed in the onset-masking condition than in the offset-masking condition in both experiments.

General Discussion

Using the visual-word paradigm, the present study investigated the effects of word onset and offset masking on the time course of non-native spoken-word recognition in noise. In two experiments, Dutch listeners heard English target words, preceded by carrier sentences that were noise-free (Experiment 1) or contained intermittent noise (Experiment 2). Target words were either onset- or offset-masked (at SNRs -6 and -12 dB) or not masked at all.

In both experiments, we observed that onset masking had more detrimental effects on spoken-word recognition than offset masking. These effects manifested as delayed fixation biases for the target objects, relative to unrelated distractors. Our findings are in line with the results reported by XXX (YYYY), who used an offline word transcription task. Similar to XXX and colleagues, we found that offset masking led to delayed target biases, albeit less dramatically than onset masking did. This suggests that both word onset and offset information contribute to spoken-word recognition in noise, with non-native listeners relying more strongly on onset information.

Based on the conclusions by McMurray and colleagues (2017), we had hypothesized that non-native listeners might adopt a similar 'wait-and-see' strategy as native listeners when faced with a degraded speech signal. While we observed some evidence in Experiment 2 supporting this notion (i.e., the presence of intermittent noise in the carrier sentences resulted in overall later target fixations), such a strategy did not affect the relative importance of onset and offset word information for non-native spoken-word recognition. Note that one important difference between the present experiments and that of McMurray et al. (2017) is that in their stimuli the whole target word was degraded rather than onset or offset parts. It is thus conceivable that listeners in their experiment delayed lexical access since they were presented with input that was masked in its entirety. In the present experiments, some part of the target word was always audible. Therefore, participants most likely tried to access lexical representations as early as possible, shortly after target word onset. When onset information was less reliable, phonological mapping was delayed resulting in a delayed target bias—sometimes reaching statistical significance only after target word offset (Experiment 2).

To conclude, our data show that when the word onset is masked, fixations to the target are delayed and that offset masking was less detrimental than onset masking. These results are in line with standard theories of spoken-word recognition that predict that lexical access occurs as early as possible.

References

Ben-David B.M., Chambers C.G., Daneman M., Pichora-Fuller M.K., Reingold E.M., & Schneider B.A. (2011). Effects of aging and noise on real-time spoken word recognition: evidence from eye movements. *J Speech Lang Hear Res, 54* (1), 243-262.

Boersma, P. (2011). Praat: doing phonetics by computer [Computer program]. *http://www. praat. org/*.

Brouwer S., & Bradlow A.R. (2016). The temporal dynamics of spoken word recognition in adverse listening conditions. *J Psycholinguist Res, 45* (5) (2016), 1151-1160.

Donnelly, S., & Verkuilen, J. (2017). Empirical logit analysis is not logistic regression. *J Mem Lang, 94*, 28-42.

Garcia Lecumberri, M. L., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Commun, 52*, 864-886.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychol., 137*(2), 151-171.

Lemhöfer K., & Broersma M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behav Res Methods, 44* (2), 325-343.

Mattys, S. L., Davis, M. H., Bradlow, A. R. & Scott, S. K. Speech recognition in adverse conditions: A review (2012). *Lang and Cogn Proc, 27*(7-8), 953-978.

McQueen J.M., & Huettig F. (2012). Changing only the probability that spoken words will be distorted changes how they are recognized. *J Acoust Soc Am, 131*(1), 509-517.

McMurray, B., Farris-Trimble, A., & Rigler, H. (2017). Waiting for lexical access: Cochlear implants or severely degraded input lead listeners to process speech less incrementally. *Cognition, 169*, 147-164.

Saslow, M. G. (1967). Effects of components of displacement-step stimuli upon latency for saccadic eye movement. *JOSA*, 57(8), 1024-1029.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *J Exp Psych: Learn Mem Cogn, 6*(2), 174-215.

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Q J Exp Psychol, 67*(6), 1176-1190.

Wood, S. N. (2017). *Generalized additive models: an introduction with R.* CRC Press.