Honors Theses at the University of Iowa

Spring 2018

Framing Expectations: The Effects of Transient Degraded Speech Conditions on Competition in Spoken Word Recognition

Victoria Shihadah
University of Iowa

FRAMING EXPECTATIONS: THE EFFECTS OF TRANSIENT DEGRADED SPEECH CONDITIONS ON COMPETITION IN SPOKEN WORD RECOGNITION

by

Victoria Shihadah

A thesis submitted in partial fulfillment of the requirements
for graduation with Honors in the Speech Pathology and Audiology

Bob McMurray
Thesis Mentor

Spring 2018

All requirements for graduation with Honors in the
Speech Pathology and Audiology have been completed.

Yu-Hsiang Wu
Speech Pathology and Audiology Honors Advisor

FRAMING EXPECTATIONS: THE EFFECTS OF TRANSIENT DEGRADED SPEECH CONDITIONS ON COMPETITION IN SPOKEN WORD RECOGNITION

By

Victoria Shihadah

A thesis submitted in partial fulfillment of the requirements for graduation with Honors in the Department of Communication Sciences and Disorders

____

Bob McMurray

Thesis Mentor

Spring 2018

All requirements for graduation with Honors in the Department of Communication Sciences and Disorders have been completed.

____

Yu-Hsiang Wu

Communication Sciences and Disorders Honors Advisor

Abstract

Speech changes continually in time. Consequently, for listeners to recognize spoken words, they must piece together the incoming message over time. As a listener hears a word like *sandwich,* they immediately activate multiple candidates from their mental lexicon with similar onsets (*sandwich, sandal, santa*). They then integrate further auditory input as it arrives, to favor or disfavor these candidates. This competition takes different forms under degraded listening conditions, for example in listeners with cochlear implants. However, it is unclear whether these differences arise from the degraded input itself, or if listeners refine this competition to adapt to poor input. Thus it was investigated how word recognition unfolds in conditions when the target word is clear, but listeners believe they are listening in noise. For the purposes of this study, a new type of noise, referred to as *framed noise*, was developed, in which a carrier sentence is presented along with background noise (e.g,. *now click on the…*), but the target word *(…ball*) is clear. This was compared to conditions of complete-noise and no noise. Lexical competition was measured using in the Visual World Paradigm, in which listeners matched a spoken word (e.g., *sandal*) to one of four pictures on the screen *(sandal, sandwich,* etc.), while fixations to each picture were recorded, revealing participants' early interpretations. We found listeners in the noise and framed condition waited for further input before fixating on the target and competitors. Listeners also activated competitors longer in the noise condition, but not in the framed. The results indicate that varying degraded auditory input will influence processing strategies more than expectation.

Introduction

The speech signal changes continually as it unfolds in real time. As a result, while listeners receive auditory input, they must accurately piece together the time varying signal to understand the intended message. However, this signal may be relayed in less than ideal conditions, such as in the presence of noise. This increases the challenges listeners confront to accurately understand speech.

In order for the listener to succeed in understanding a sentence, they first must access the meanings of the individual words in the mental lexicon. Research suggests that as normal hearing (NH) listeners receive acoustic speech information in quiet, possible lexical candidates are immediately activated and evaluated in their mental lexicon (Marslen-Wilson & Welsh, 1987; McQueen, 2008, for a review). The set of candidates is continually narrowed down in real time as more information arrives (Marslen-Wilson, 1987; Allopenna et al., 1998; Reinisch, Jesse, & McQueen, 2010). For example, as the listener begins to hear *candy* (e,g., when they've heard only the initial information *ca…*), they immediately activate possible candidates like *candy*, *candle* and *captain*. Later on, as they receive further auditory input, such as the /n/, activation is suppressed for words like *captain* that are not consistent with the additional phonemic information.

In the present study, we ask how these dynamic competition mechanisms differ when people expect challenging listening conditions. We start by exploring the general properties of lexical access in quiet and how they are measured. We then discuss how lexical access differs in noise, before presenting an experiment which uses a novel type of noise to manipulate listeners' expectation of a noisy signal.

Cognitive Mechanisms in Spoken Word Recognition

There are thought to be four cognitive mechanisms used in this temporal processing strategy: *immediacy, parallelism, incrementality*, and *competition*.

The moment a listener begins to receive spoken input, words in the mental lexicon are immediately activated. This is the principle of *immediacy* (Marslen-Wilson, 1987; McClelland & Elman, 1986; Norris & McQueen, 2008). For example, when hearing the word *beaker*, listeners will momentarily fixate on a picture of a *beetle* (Allopenna, Magnuson, & Tanenhaus, 1998), implying that it is being considered at the semantic level. Immediacy allows the listener to understand the intended output before they have perceived it entirely. Since lexical access begins from this initial information, this creates temporary ambiguity as there is an exhaustive list of possible candidates when only a few phonemes have been heard . As a result of this temporary ambiguity, multiple words are activated simultaneously*,* the principle of *parallelism,* (Allopenna, Magnuson, & Tanenhaus, 1998; Marslen-Wilson & Zwitserlood, 1989). Activating candidates in parallel allows listeners to efficiently consider multiple candidates for the target simultaneously. That is, not only do listeners activate the target candidate, they also simultaneously activate other candidates with similar phonemic onset information (Marslen-Wilson, 1987; Marslen-Wilson & Zwitserlood, 1989). For example, when the /k/ in *candy* is heard, the target and competitors (e.g. *candle*, *castle*, etc.) will be activated.

In running speech, the evidence in the signal for a given word is inherently probabilistic. Consequently as more auditory cues are received, listeners must continually adjust the degree to which they favor or disfavor candidates (Marslen-Wilson & Welsh, 1987; Frauenfelder, Scholten, & Content, 2001). This ongoing process incorporates information in real time and is known as *incrementality*. For example, after listeners activate a set of candidates with similar onsets, subsequent (medial and offset) phonemic information will prompt them to favor and disfavor those candidates, keeping candidates that are consistent with the phonemic information (Marslen-Wilson, 1987; Marslen-Wilson & Zwitserlood, 1989).

Finally, as the words are activated, they dynamically *compete*. As the listener receives more auditory input favoring a target word, it inhibits words that are less likely (Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Luce & Pisoni, 1998). For example, when the /z/ in *wizard* is heard, it creates additional activity for the target word, and *wizard* can then inhibit competitors (e.g. *whistle*) to suppress their activation even faster (Marslen-Wilson, 1987; Marslen-Wilson & Zwitserlood, 1989). This competition is key for narrowing down the set of possible competitors until there is only one left, the target word.

These four cognitive mechanisms—*immediacy*, *parallelism*, *incrementality*, and *competition*—are posited to describe processing in typical conditions. However, by tuning how they work, they may also offer the ability to effectively adjust expectations and processing strategies throughout running speech in order to accurately understand the intended message.

Visual World Paradigm

This dynamics of this competition process can be measured with the Visual World Paradigm (VWP) (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Allopenna et al., 1998). The VWP tracks participants' eye movements to pictures on a screen as they respond to spoken instruction directing them to one of the on-screen pictures. Each picture corresponds to potential lexical candidates such as the target (e.g., *sandal*), a competitor with a similar onset called the cohort (e.g., *sandwich*), a rhyme competitor that overlaps at offset (e.g., *candle*), or an unrelated object that is not phonologically related to the target (e.g., *parrot*).

Fixations to each picture are recorded in real time throughout this task. In order for a listener to select a picture on the screen, they must look around and fixate on the various pictures as they hear the auditory input in real-time. Since listeners can make 3-4 fixations between the word and the response, these fixations can reveal how the participant may have interpreted and processed the stimuli at each moment in time. The VWP provides a millisecond-by-millisecond measure of how each option (e.g., in this example, the target, cohort, rhyme, unrelated) is considered during real-time word recognition. While the final choice (e.g., what they click on) reflects the listener's final decision, these ongoing fixations reveal what items are considered on the way to that final decision. The VWP provides researchers a means to visualize and analyze the timecourse of lexical processing.

Using this paradigm Allopenna et al. (1998; see also Ben-David et al., 2011) first demonstrated that at word onset, listeners fixate on both the target and possible cohort competitors. As more auditory information arrives, the target object receives increasing fixations (as it matches the incoming auditory input), while cohort competitors decrease in fixations. This pattern is due to incoming phonemic information not consistently matching all of the cohort competitors and therefore lexically unlikely to be said in that point in time. The onset of a word is not the only valuable auditory information that affects lexical activation. Rhyme affects are found in VWP (Allopenna et al., 1998): while hearing a target, listeners activate words that rhyme with it to ensure all possible words are considered even when onset information is initially mismatched. For example, when *beaker* is heard, *speaker* may be activated as well. However, rhyme competitors are not as strongly activate as onset competitors (cohorts), which exert a larger influence on word recognition (Allopenna et al., 1998). Lastly, unrelated words with no rhyme or phoneme onset similarities are barely if at all activated in the mental lexicon as the listener receives auditory input.

Thus, the VWP can measure the dynamics of word recognition and shows a pattern of fixations that is consistent with the principles described earlier. How effectively these dynamics work in lexical processing may depend in part on contextual variables such as background noise.

Role of Competition in Degraded Speech Conditions

When speech is heard in the presence of background noise, perceptual ambiguity is expected to arise for the NH listener, as the bottom up information is less clear and therefore should be considered less trustworthy. Listening to speech in noise is difficult, but significantly more so for those with assistive listening devices such as Cochlear Implants (CI) (Muller-Deiler Schmidt, Rudert, 1995; Nelson & Jin 2003). However, even in quiet, CI users and others with assistive listening devices must manage a consistently degraded input due to hearing loss and the limitations of their technology. These limitations in perception contribute to the wide variability of speech recognition in those with assistive listening devices (Fu & Nogaki, 2004). Yet, despite these longstanding clinical problems, the range of cognitive processes individuals use when hearing speech in noise is not entirely understood. Unlocking the underlying mechanisms of spoken word recognition in degraded conditions, might lead to better, more effective support to those with hearing loss and various assistive listening devices.

The question we ask here is how listeners adjust their dynamics of lexical processing to accurately piece together incoming speech information when they are in uncertain or difficult listening conditions? Three recent studies have examined this issue and converged on a similar pattern of lexical competition in noise.

Brouwer and Bradlow (2015) first used the VWP to evaluate the temporal dynamics of lexical competition in noise and background speech. Participants heard target words either in quiet, mixed with broadband noise, or mixed with background speech. Participants' eye movements were tracked to measure competition between candidates including the target, cohort and rhyme. Brouwer and Bradlow (2015) found that the participants in the degraded conditions displayed a different profile of competition than the typical ones found in quiet (Allopenna et al., 1998). Relative to the quiet condition, participants showed a delay in fixating to the target word, with increased looks to the onset competitor, and a smaller increase in rhyme fixations. This suggests an *increasing competition* strategy in which under adverse listening conditions, listeners deliberately compensate for acoustic uncertainty by holding off on a final decision, waiting for further input. They do this by maintaining competition between likely competitors (e.g. looking at both *beaker* and *beaver* longer before deciding) until they can confirm they heard the target, avoiding incorrect responses.

Ben-David, et al., (2014) evaluated age-related differences in spoken word recognition in quiet and noise in both younger and older adults. Listeners followed spoken instructions that referred to objects on the screen (e.g. *look at the candle.*). Their data revealed similar processing patterns in younger and older adults in the majority of conditions. However, there were age-related differences regarding rhymes in the noise condition. Specifically, participants delayed fixating only on the target when there was a rhyme competitor in the noise condition. This competitor effect is similar to Brouwer and Bradlow's (2015) *increasing competition* strategy they found. When participants from both studies were in uncertain, degraded speech conditions, they kept competitor available by maintaining longer activation of them until further input confirmed their target choice.

Farris-Trimble, et al. (2013) found similar results to Brouwer and Bradlow (2015). They used the VWP to examine three groups: adult prelingually deafened cochlear implant users, NH adults who listened with 8-channel CI simulation, and a NH control group. CI users also face a degraded or uncertain input, though the exact form of this is quite different to general speech in noise. However, both CI users and NH listeners are regularly faced with uncertain conditions. Results indicated that the CI and NH participants with simulation had similarities and differences in their processing strategies from NH listeners with a clear speech signal. Both CI users and the CI simulated group showed decreased cohort fixations and increased rhyme fixation.

Meanwhile, CI participants delayed fixating to the target object the simulated group had less delay (though they were still delayed relative to NH listeners with unsimulated speech). Moreover, while CI users maintained activation of competitors (e.g. using the *increasing competition* strategy), CI simulation listeners fully inactivated the cohort competitors. This difference between CI users and NH listeners in simulation may indicate that CI users learn to adapt more readily their usual listening strategies due to long-term degraded input from their hearing loss, while those NH listeners who are only briefly in a simulation, are unable to develop this processing strategy as well. This difference also highlights that listening in noise involves dynamic lexical processing strategies that quiet does not require.

However, this increasing competition strategy is not the only one available to listeners. There is another possible, contrasting strategy. Farris-Trimble, McMurray and Rigler (2017) found substantially different results from previous studies on lexical processing in degraded speech. They examined prelingually deafened CI users, and NH listeners in severely degraded conditions (4-channel CI simulation; for comparison the prior study used 8-channel), and a NH control group. Both the CI and NH groups with simulation showed a substantial delay in activating the target word (about 250 msec). As a result of this, they showed less competition from the onset competitors and an increase in competition from the rhyme competitors. In highly degraded conditions, the two groups held off on accessing the lexicon, and this reduced competition from onsets (since more of the word had arrived). This suggested a strategy in which participants collect incoming auditory input, without accessing the mental lexicon immediately for any specific competitor in order to be more careful due to the degraded input. Only when substantial input is accumulated, do they access the mental lexicon and proceed with activating and deactivating competitors. This strategy is substantially different compared to the *increasing competition* strategy, and can be loosely termed *wait and see*.

This strategy was used similarly by both populations of participants (CI users and simulated NH) in degraded conditions, indicating that listeners in general (not just CI users) may utilize significantly different lexical processing when faced with extremely degraded conditions,. Rather than selecting a single competitor as soon as enough auditory input is received, listeners may hesitate, accumulating more input before accessing the mental lexicon.

It is clear that adverse listening conditions influence the strategy listeners use to identify the target word. Listeners can either activate multiple competitors and maintain them for a longer amount of time (*increasing competition*) or they can hold off on activating on any lexical items until receiving further input (*wait and see*). However, it is not clear whether the altered lexical processing strategies are due to the degradation of perceptual cues or due to the expectation that there will be degraded perceptual cues, altering strategies ahead of time. Is degraded input simply too difficult to understand using typical processing strategies, resulting in an increased competition? Or, did the context of the preceding degraded conditions lead listeners to expect that similar strategies would be needed f allowing them to adjust beforehand? The studies do not draw a clear line between which strategies are used when faced with degraded input and when degraded input is expected beforehand.

McQueen and Huettig (2012) attempted to address this issue more directly. They used the VWP to examine spoken word recognition when the participants are led to believe the target word is degraded and it was in fact not. This was accomplished by replacing various phonemes with noise in the carrier sentence leading up and following the target word (but leaving the target word undistorted). In the baseline (undistorted) condition, participants fixated on cohorts longer than other possible choices. However, when the stimulus leading up to the target became degraded and therefore uncertain, participants altered their strategy: the participants looked less at cohort competitors and more at rhyme-competitors. McQueen and Huettig (2012) described this flexible processing as due to participants anticipating noise on the target word, but not receiving it. McQueen and Huettig (2012) hypothesizes that this change in processing dynamics indicates that the overall expectation of uncertainty leading up to the target word altered how the participants decided to process the incoming target auditory information. This suggests that there may be a unique role for higher level expectations.

However, there are several concerns with this study that prevents a thorough understanding of these mechanisms. First, participants' expectations for a noisy target were never fulfilled. There was not a condition in which listeners anticipated a distorted target and that anticipation proved accurate. Instead, they only heard conditions in which they anticipated a distorted target, and this was violated by a clear target. This additional condition would have provided data to evaluate if their strategies altered or interacted when their expectations were proven right or wrong.

Secondly, participants were not given the target word as a choice on their screen. Therefore, they were never able to fixate on the correct response, but only cohort competitors or unrelated competitors. The patterns that were found from these competitor fixations seemed to indicate the *wait and see* approach. However, without recorded target fixations to compare to the competitors, it is unclear whether they were waiting at all. Further, the lack of a target word on the screen raises the possibility that the fixation patterns in this study may also have reflected compensatory or secondary processes. Thus, the data prevents a thorough understanding of spoken word recognition in uncertain conditions leading up to the word, because they did not have the option of choosing the target on the screen of visually recognizing the word like they did for the competitors. The process of recognizing the target word in the uncertain and degraded conditions is therefore unclear.

The Current Study

The present study set out to investigate the real-time competition mechanisms used in lexical processing by independently manipulating the expectation of a degraded input and the actual degraded input. We asked whether the processing strategies used in uncertain and degraded conditions are responses to poor auditory input or at least partially reflect higher level expectations.

The stimuli delivered in each condition was a semantically neutral carrier sentence followed by the target word. The carrier sentence was used to influence participants' expectations of the level of degradation present on the following target word (as in McQueen and Huettig, 2012). Participants identified the target's referent on the screen from a display containing four pictures: target, cohort, unrelated and unrelated. To create a condition in which the carrier sentence was noisy, but the target was clear, we developed a new type of noise, *framed noise* (see Figure 1). In this condition, the carrier sentence is delivered in background noise, but the target word *(…ball*) is clear below 4 kHz, preserving the key perceptual information needed to accurately recognize the target word. The 4 kHz ceiling provides a constant canopy of noise that preserves an overall expectation of noise even when hearing the clearly audible target word initially. This condition separates the expectation of noise and the actual perception of the target word.

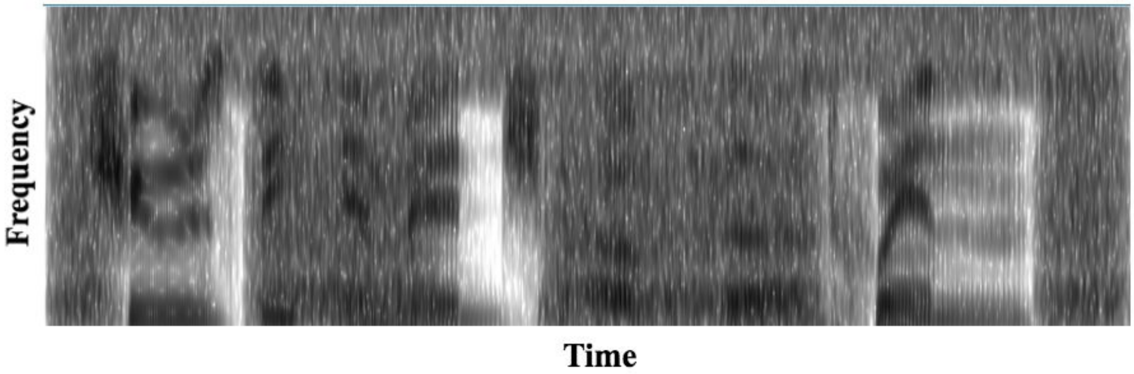

Figure 1. *Spectrogram of Framed Noise.* Example of framed noise condition with sentence: "*please click on the twin."* White noise delivered simultaneously with the carrier sentence, but not during the target word. There was a

consistent canopy of noise at 4 kHz and above but the target word was perceptually clear below 4 kHz. Gaps in the noise were added throughout the carrier sentences in the framed (and noise) condition to decrease the jarring affect a noisy carrier sentence leading into a perceptually clean target word may create.

Second, in the *noise* condition, listeners heard the initial carrier sentence and the subsequent target word in speech shaped noise. This maintained a connection between expectation of noise throughout the stimulus and actual perception of the noise on the target. This condition was anticipated to provide similar results found in previous research on speech in noise (Farris-Trimble et. al., 2013 and 2017; Brouwer & Bradlow, 2015; McQueen & Huettig, 2012).

Lastly, in the *quiet* condition, listeners heard the initial carrier sentence and subsequent target word in quiet. This condition maintained a connection between expectation of quiet throughout the stimulus and actual perception of the target word in quiet. This condition is considered a baseline condition in which the dynamics of lexical activation should look like typical results found in spoken word recognition in quiet (Allopenna et al., 1998; Ben-David et al., 2011).

The dynamics of lexical activation in each condition were monitored with a VWP task. We hypothesized that if lexical processing is not shaped by have high-level expectations for noise, the framed condition should look like the quiet condition since both conditions provide a target word that is in quiet with no degradation of the cues. However, if listeners do have high level expectation strategies, results in the framed condition should look like the noise condition as both create an expectation that there will be a degraded target word due to noise in the carrier sentence.

Methods

Participants

Eighteen NH adults participated in this study. Participants were recruited from the departmental subject pool for the Department of Psychological and Brain Sciences at the University of Iowa. All participants self-reported as normal hearing and native monolingual English speakers. One participant was excluded from this study as they did not complete all of the required trials and one participant was excluded due to growing up in a bilingual household. This left 15 participants' data in the study. All participants underwent an IRB approved informed consent procedure.

Design

On every trial, the participant heard a carrier sentence instructing them to click on a single word. The array of options on the screen included four pictures: the target, a cohort, and two unrelated words. The Target and Cohort were paired with one another, ensuring there was initial phonemic overlap, but no semantic overlap. Furthermore, targets and cohorts also had similar syllable structure (e.g. *captive* and *captain* were paired). There were a total of 30 unique target-cohort pairs. Each target-cohort pair was combined with another pair that was phonologically and semantically unrelated (e.g., *captain/captive* with *beaker/beetle*). Thus, this second pair served as the unrelated controls for the first. Pairings were checked to ensure there was no phonemic or semantic overlap between the two target-cohort pairs. We randomly grouped two target-cohort pairs into sets of four items to create 15 item sets. This was done three times to create three different randomization lists (see lists in Appendix A). There was no overlap in sets between each list. Each participant was run with one of the three unique lists. Each set of four in the list rotated in which item was the target word. Each target word was delivered nine times divided evenly across the three conditions. In total, there were 540 trials for each participant. There were sixteen carrier sentences rotated nearly equally for every participant. Multiple carrier sentences were used to ensure no semantic or prosodic information would influence or bias the participants' fixation patterns.

Stimuli

Auditory Stimuli. Auditory stimuli were recorded by a male native English speaker with a Midwestern dialect in a sound attenuated room. Recordings were made using a Kay CSL 4300B A/D board at 44100 Hz. All recordings underwent noise reduction in Audacity by extracting a pure noise profile from a silent portion in the audio recording. The profile was then applied to the entire recording, reducing any noise that matched that profile.

Each target word was recorded with the same neutral carrier phrase (*Please choose the picture of the **X***), followed by a pause and then the target word. This was to ensure consistent and even prosody across all words. Multiple exemplars of each target word were recorded, with the best three selected. The target word was excised from the original recording to be spliced on to the carrier sentences. Target words were cut at the zero crossing closest to the onset and offset of articulation. No additional silence was added to the target word onset.

The sixteen carrier sentences were all recorded with the same word (*dud*) at the end of the sentence, to ensure equal pausing and sentence intonation patterns (Appendix B). The best four exemplars of each sentence were chosen and excised from the original recording, and cut from the filler target word (*dud*) at the nearest zero crossing at the end of the sentence.

Finally, the carrier sentence and target words were spliced together in all possible combinations using a MATLAB script. There was no overlap or added time between the carrier and word. Stimuli were amplitude normalized using Praat. An additional 100 msec was added to the beginning and end of the entire stimulus.

In the noise and framed noise conditions background noise was stimulus shaped noise. The noise was generated by reading in all exemplars of the target and carrier sentence. We then extracted the long term average spectrum for all of the stimuli. Next, we generated white noise. Finally, the noise was filtered by LTAS so the spectrum of the noise matched the most common frequencies of the stimuli. We then introduced 1-4 randomly distributed short gaps into the carrier portion of the noise. Gaps also were framed with noise above 4000 khz. These randomized gaps in both noise and framed noise ensured participants would not be “surprised” to suddenly encounter a clear target word, as the gaps created the expectation that random segments (in both the carrier and target word) could be noise-free.

In the noise condition, this noise was mixed throughout the entire carrier sentence and target at a ratio of 0.8 signal and 0.2 noise. This ratio was determined to provide enough noise to create relatively difficult tasks compared to the quiet condition, without giving the participant too little bottom up perceptual information to perform well in the task. In the framed the same noise (at the same ratio) was mixed into the carrier sentence, but was for the target word, it was limited to 4 kHz and above.

In the quiet condition, we low pass filtered the stimuli (< 4 kHz) to ensure the framed condition and the quiet condition were equal in key perceptual information on the target.

Visual Stimuli. Visual stimuli consisted of clip art images created through a standard lab procedure (McMurray, Samelson, Lee & Tomblin, 2010). For each word, multiple images were downloaded from a commercial clip art library. A focus group of graduate and undergraduate students selected the best representative of the word from the group of images. Additional selective editing was done to images when necessary to ensure precise screen dimensions were met. Official approval was done by lab members with extensive knowledge and experience with the VWP.

Procedure

Participants underwent a verbal consent procedure, and filled out a short demographic questionnaire. Next, the participant was seated in front of the computer screen while the head-mounted eye-tracker was set-up and calibrated. After that, the participant received instructions (both verbally and in written form) on the tasks and was given the opportunity to ask questions. Finally, they underwent a practice drift correction, and began the experiment.

On each trial, the participant saw four pictures on a 17'' (1280 x 1024 pixel) computer monitor. The four pictures were 300 x 300 pixels and located, 50 pixels from the edges of the monitor. At the beginning of the trial, the four pictures were displayed with a red circle in the center for 500 msec. This gave the participant the opportunity to familiarize themselves with the visual stimuli and their locations. This 500 msec pre-scan period minimized eye-movements due to visual search after the auditory stimulus was heard. After 500 msec, the red dot turned blue, indicating that the participant could select it with a computer mouse to start the trial. Once they clicked on the circle, the participant heard a neutral carrier sentence (*please click on the…)* followed by a target word (*...ball).* After hearing the auditory stimulus, the participant clicked on the picture they believed matched the target word. This ended the trial. Stimuli were delivered over supra-aural headphones, specifically amplified by a Sennheiser, HD 201.

Eye-tracking recording and analysis

Eye movements were recorded with a head mounted SR Research Eyelink II eye-tracker. Both corneal reflection and pupil in both eyes were tracked whenever possible (though only the data from the eye with the best calibration was used). A standard 9-point calibration was used. Periodically, (every 20 trials), a drift correct procedure was performed take into account natural drift over time and maintain calibration. If the participant failed the procedure, the researcher recalibrated the eye-tracker. Point of gaze was sampled and recorded at every 4 msec, starting at the beginning of each trial and continuing until a picture was selected by the participant.

The recorded eye-movements from each trial were automatically grouped into saccades, fixations and blinks using default parameters. Each saccade and the subsequent fixation was grouped into a “look.” The look began at the onset of the saccade and ended at the offset of the fixation in each trial. In identifying what object the participant fixated on, boundaries of the objects in the screen ports were extended horizontally and vertically by 100 pixels. This accounted for any noise in the eye-tracker and did not create any overlap between object images.

Results

Accuracy

We began with an analysis of accuracy between conditions. Overall accuracy in the quiet condition was 99.44%, noise was 92.42%, and framed was 99.77%. To confirm these percentages were significant, we used an ANOVA. As a result, we found a significant difference in accuracy, $F(2, 30) = 283.113$, $p < .01$, indicating conditions did affect participants' accuracy. To further analyze these differences, we used a paired sample t-test, finding a significant difference between quiet and noise, $t(15) = 18.393$, $p = < .01$. This indicates that participants were substantially affected in their accuracy when listening in noise. Furthermore, no significant difference was found between quiet and framed, $t(15) = -1.640$, $p = .122$, showing that accuracy was not significantly affected when participants anticipated noise in the framed. Lastly, there was a significant difference between the framed and noise condition, $t(15) = 16.962$, $p = < .01$, indicating that expectation (framed) did not affect accuracy like degraded input (noise). These differing results in accuracy between conditions indicate that even when incorrect trials are excluded, we are pulling from robust data that will provide valuable results.

Eye-movement analysis

For analysis of the fixations, the trials in which the subject did not select the correct target on the screen were removed before analysis

We first analyzed the pattern of proportion of fixations to each of the classes of objects beginning from trial onset (0 msec in the figures) to the offset. Participants did not hear significant phonemic target information until several hundred msec in. Figure 2 and 3 show the proportion of fixations for the target and the cohort competitor as a function of time and condition. These figures suggest that the noise condition (degraded or not) influenced the initial proportion of fixations for both the target and cohort even before receiving any key onset information. Even before being the target word (0-300 msec), participants in quiet showed increased proportion of fixations to all four objects. This may be due to participants being confident in their ability to accurately identify the target due

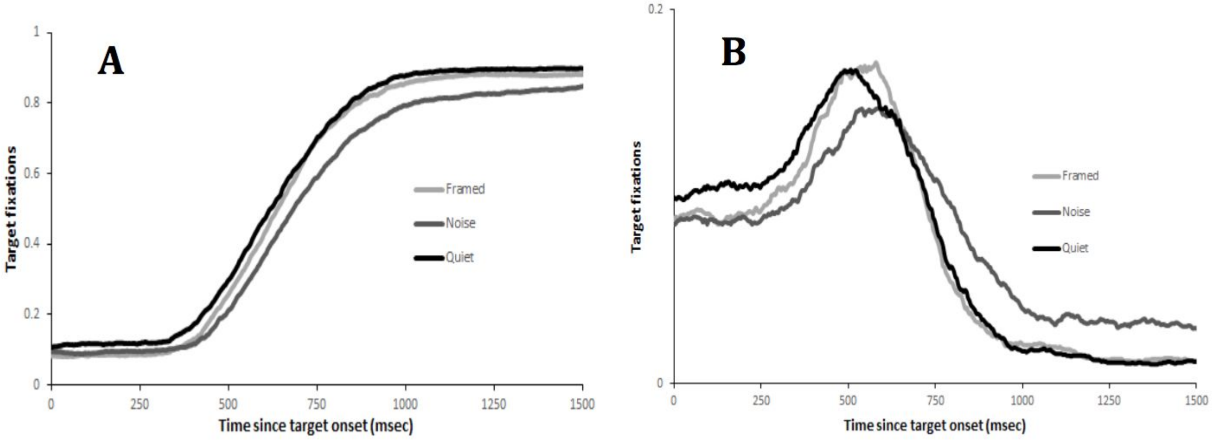

Figure 2. (A) *Time Course of Fixations to Target Across Conditions.* Proportion of looks to the target between conditions. (B) *Time Course of Fixations to Cohort Across Conditions.* Proportion of looks to the cohort between conditions.

In quiet, as participants received phonemic information of the target word, they increased fixations to both the target and cohort faster than in the other conditions. This may be attributed to overall clean auditory input and confidence that they know the word before hearing it in its entirety. Additionally, compared to quiet, both the framed and noise condition have delays in target and cohort fixations. Furthermore, the framed condition has earlier activation of both the target and cohort compared to the noise condition, but was somewhat delayed relative to the quiet condition. This suggests that although participants were initially uncertain of the input, the dynamics of spoken word recognition rapidly shifted as they perceived a clean target to reflect the pattern seen in quiet.

Lastly, after the initial activation of the target and cohort (around 250 msec), the quiet and framed conditions show similar patterns of fixations, with similar-looking slopes and asymptotic proportion of fixations. However, noise has a differing pattern, with an overall more delayed and decreased fixation pattern for both the target and cohort. This may indicate that the participants were more affected by conditions that significantly degraded the quality of the input, than by conditions that only led them to expect such degradation.

Target Analysis. To quantify the differences found above we used a nonlinear curve fitting technique. Target fixations were fit to a four parameter logistic (Farris-Trimble, McMurray & Rigler, 2015), estimating the shape of the timecourse of looking to each competitor for each participant. The *crossover* point (msec) describes the overall delay or shift of the curve in time. The *slope* (the derivative of the function at the crossover) reflects the speed at which activation builds. The upper asymptote is the degree of final fixations, relating to the confidence of the participant in their final decision (McMurray et. al., 2010). The lower asymptote is the lowest fixations to the target. We also computed an additional, new measure: the 25% crossover point. This was defined as the point which the curve crossed 25% of the way between the lower and upper asymptotes, and provides information on the early fixation patterns during the onset of the target word. We did not analyze onset baseline, crossover, and slope for target fixation data.

This function was fit to each participant's data in each condition using a constrained gradient descent technique (McMurray, 2017). Subsequently, we conducted a repeated measures ANOVA to analyze the effect of noise condition on select parameters relevant to our study. The quality of fits were calculated by finding the average. The quality of the curve fits was calculated by averaging all the correlated values from each correct trial, R=0.996717. The minimum correlation was 0.958748. The high R value indicates that the fits found were highly representative of the pattern of fixations for each condition. We first analyzed the effect of noise condition on the curve (fixation pattern) on the 25% crossover point. The results showed a marginally significant main effect of condition, $F(2,30) = 12.204$, $p < .054$. Paired-samples t-tests found no significant difference between quiet and framed, $t(15) = .381$, $p = .709$, with similar crossover points between these conditions. However, there were significant differences between quiet and noise, $t(15)= 2.355$, $p = .033$ and framed and noise, $t(15) = 2.526$, $p = .023$. Participants significantly slowed fixations to the target in the noise condition compared to quiet and framed, indicating onset perceptual information may play a role in fixation patterns at 25%.

Next, we examined peak looks to target. Results indicated significant differences between groups, $F(2,30) = 12.204$, $p < .001$. We thus conducted paired-samples t-tests and found there was no difference between the quiet and framed noise conditions, $t(15) = 1.239$, $p = .234$. However, there was a significant difference between the quiet and noise conditions, $t(15) = 4.044$, $p < .001$ and between the framed and noise condition, $t(15) = 4.745$, $p < .001$. The significant difference between conditions indicates that degradation of phonemic information decreases ultimate looks to target, but a mismatch in expectation does not.

Cohort Analysis. To analyze fixation patterns between conditions for cohort competitors. Cohort fixations were fit to a double Gaussian function. This function has six parameters: *Mu* reflects the time at which the function reaches its peak; the peak height (*h*) represents the overall maximum looks to cohort; *b1* and *s1* reflect the initial base and slope while *b2* and *s2* reflect offset base and slope. The quality of the curve fits found with the Gaussian function was calculated by averaging all the correlated values from each correct trial, *R* = 0.984047; the lowest correlation number was 0.905517. The high average R value indicates that the fits were highly representative of the pattern of fixations for each condition.

A repeated measures ANOVA revealed a marginally significant main effect in *mu* between conditions, *F*(2, 30) = 2.737, *p* < .081. Further analysis with a paired t-test showed a significant difference between the quiet and framed conditions, *t*(15) = -2.513, *p* =.024, indicating that peak cohort fixations were earlier for quiet than for the framed condition. However, there was no significant difference between quiet and noise, *t*(15) = -1.912, *p* = .075 and framed and noise, *t*(15) = -.031, *p* = .976.

Next, we analyzed the initial baseline for cohort fixations and found no significant main effect, *F*(2, 30) = .079, *p* = .924. These results suggest the expectation of noise did not cause participants to initially suppress their looks to competitors.

Next height was examined with a repeated measures ANOVA, yielding a significant main effect, *F*(2, 30) = 4.140, *p* = .026. A paired t-test analyzed these differences further: there was a significant difference between quiet and noise, *t*(15) = 2.460, *p* = .027, indicating degraded phonemic information *decrease* fixations to cohort competitors. There were no significant difference between quiet and framed, *t*(15) =.056, *p* = .956, indicating that the initial mismatch in the framed condition did not significantly impact how much participants fixated on the cohort. However, significant difference was found between framed and noise, *t*(15) = 2.454, *p* = .027. The difference in height of cohort fixations indicates that degraded input influenced fixation patterns.

Finally, we analyzed *offset slope* for cohorts. A similar repeated measures ANOVA found a marginally significant main effects, *F*(2, 30) = 4.58, *p* = .081. Further analysis with a paired t-test yielded no significant difference between quiet and framed and the quiet and noise conditions, *t*(15) = 2.02, *p* = .062, *t*(15) = -1.15, *p* = .268. However, there was a significant difference between the framed and noise conditions, *t*(15) = -2.949, *p* = .010, indicating participants more quickly suppressed fixations on the cohort in framed noise than in pure noise. This behavior may be due to the perceptually clean input in framed compared to the degraded input in noise.

Discussion

Summary

This study isolated the role of expectations for degraded input, from the quality of the input. We manipulated listeners' expectations for degraded speech independently of the level of degradation in the actual target word. This allowed us to ask whether the processing strategies used in uncertain and degraded conditions are responses to poor auditory input or reflect higher level expectation-based mechanisms. We used the VWP to evaluate fixation patterns in NH participants when identifying words in quiet, noise, and framed noise. In the latter condition, the preceding carrier sentence was used to create the expectation of noise, while the target word itself was perceptually clear.

The results indicated substantially different fixation patterns in noise compared to quiet. Fixations to the target and the timing of peak cohort fixations were slowed and overall reduced in quiet relative to noise. In the framed condition, there was an initial delay in fixation of target and cohort competitors similar to the patterns found in noise. However, peak looks to target were higher and decreased sooner in framed than in noise. However, statistical analysis could not confirm this initial suppression of target or cohort in any of the conditions and did not show strong differences between framed and quiet conditions.

The overall proportion of fixations were less due to expecting subsequent degraded input and more a result of actually receiving degraded input. This suggests that the level of degradation in the target word was the main determiner of changes in lexical activation, specifically *delays* in the pattern of fixation. For example, at the onset of the target word in framed and noise, participants delayed activating fixations to the target and cohort.

Limitations of the Current Study

Two limitations of this study are worth exploring before we discuss the theoretical implications. First, this study had a small sample size due to time constraints: 17 participants were tested, and only data from 15 were analyzed. As a result, there may have been differences in the data that went undetected due to the small sample size. For example, the 25% crossover point in target fixations, numerically reflected the predicted difference between quiet and framed noise, but this did not reach significance. If additional participants were in the study, perhaps these data would have showed significant differences, providing further information on the lexical processing strategies in spoken word recognition.

Another limitation of the study is that only the trials where the participant chose the correct target word were analyzed. The incorrect trials – which would have had largely different fixation patterns – were not included in the analysis. This was intended to allow the fixations to reflect a clear representation of the competition mechanisms at play when lexical access was ultimately successful. This may have affected the strength of competition we evaluated. This could have altered the overall pattern of fixations included in the study. However, incorrect trials may reflect a variety of additional factors that were not of interest here: participants getting distracted, falling asleep, not putting effort in, etc. Therefore, excluding incorrect trials was likely the most reflective of spoken word recognition.

The Effects of Spoken Word Recognition in Noise

As we described earlier, existing work supports two broad conceptualizations of the dynamics of lexical access in noise: *increasing competition* and *wait-and-see*. Increasing competition posits that listeners adapt their lexical access mechanisms to poor listening conditions by activating competitors for a longer period of time, ensuring their choice is in fact the right one (Brouwer & Bradlow, 2015; Ben-David, et al., 2014; Farris-Trimble, et al., 2013; McQueen & Huettig, 2012). In contrast, the wait-and-see approach posits that individuals adapt by waiting to activate any words (both target and competitor) until enough phonemic information is received to identify the correct one (Farris-Trimble, McMurray & Rigler, 2017).

In the present study, participants were significantly delayed in both target and cohort activation in the noise condition, indicating they altered their processing mechanisms when facing degraded input. There was no pattern that indicated initial additional activation of competitors in tandem with the onset of the target (increasing competition strategy) (Brouwer & Bradlow, 2015; Ben-David, et al., 2014; Farris-Trimble, et al., 2013). However, at the offset of the target word the results indicated substantial activation of the cohort competitor while in other conditions the competition was concluded sooner. This reflects an increasing competition strategy.

At the same time, the initial activation pattern seems to mirror wait and see (Farris-Trimble & McMurray, 2017; McQueen & Huettig, 2012). Farris-Trimble, McMurray and Rigler's (2017) findings revealed a wait and see approach much like what we found in the current study. In their study, the NH and CI users in 4-channel simulation showed a substantial delay before fixating the target or its competitors, and as a result showed *reduced* fixations to onset competitors. We show the same thing: listeners were delayed to fixate the target and showed reduced peak cohort fixations. Their study and the current one both found similar wait and see strategies at least initially in the degraded conditions. Ours extends this by showing that in conditions under which listeners expected noise and did not receive it, they did not adopt this strategy. Therefore, these dynamic patterns in noise indicates an interaction or hybrid of strategies similar to both increasing competition and wait and see.

Farris-Trimble, McMurray and Rigler (2017) found that in extremely degraded input and listening to words in isolation, CI users and NH participants were utilizing what seemed to be the *wait and see* approach. However, the current experiment employed much less degraded input and target words were imbedded in a sentence. Meanwhile, we still found results similar to the *wait and see* approach. How then are we finding related competition strategies in substantially different degraded conditions? Perhaps the studies reflect that listeners use *wait and see* or a variation of it in degraded conditions no matter the level of degradation, indicating that this strategy is more common and ongoing of a process in both NH listeners and CI users.

Furthermore, the results indicate that listeners are easily negatively impacted by degraded conditions in everyday environments.

The Effects of Spoken Word Recognition in Framed Noise

The current study found that participants in the framed condition reduced target onset fixations to both the target and cohort. This pattern of fixations *during the carrier sentence* mirrored those found in the noise condition. However, the delay in target onset fixations in framed was slightly less when compared to noise. In addition, after the onset of the target word, the slope of target and cohort fixations in the framed condition is steeper and initiated sooner than in noise, suggesting a rapid change toward the pattern found in the quiet condition. The statistical findings showed quiet and framed did not differ, while the fixations in the noise and framed conditions did. These findings indicate that expectation alone played a minimal role in how participants processed speech.

The lack of any significant difference between framed and quiet may have derived from listeners more heavily relying on phonemic information rather than their expectations of the subsequent information. If this is the case, expectation alone did not play a role. The unique pattern of fixation was not due to higher level expectation but rather the result of managing differing input.

McQueen and Huettig (2012) findings furthered the pattern conceptualized as the wait and see strategy when listeners expect noise on the target. However, the key difference between the current study and McQueen and Huettig (2012) is their findings were the result of participants being lead to expect noise, but never actually having the noise on the target word. In

contrast the current study provided both the expectation of degraded input and actual degraded input. These contrasting results may be an indicator of several things. First, the level of noise in our experiment was possibly too low, hinting at a framing condition, but not actually delivering a substantial difference for participants to adjust to. Second, it is possible that since McQueen and Huettig (2012) did not provide a target for participants to fixate on, they began processing the input in an atypical, metacognitive way. Thus the mismatch between our studies is more due to an extraneous variable in previous research rather than a unique processing strategy.
However, due to Huettig and McQueen (2012) not providing targets on the screen for participants to fixate on, the actual pattern of wait and see in their conditions remains uncertain. Our results indicate a possible relationship between noise and the wait and see approach, however.

In summary, the current study showed that when participants are faced with degraded input, they will utilize a variation of both the wait and see, and increasing competitions approach. Participants both reduced initial activation of competitors, gathering more phonemic information, and delayed peak cohort activation and subsequent reduction of it. Furthermore, when participants expected noise on the target and did not receive it, their accuracy and fixation patterns did not alter significantly. Framed noise therefore did not mirror that of noise, indicating that perceptual information is more valuable or heavily weighted when adjusting the lexical processing strategies engaged in spoken word recognition.

Implications for Audiology and Cochlear Implant Users

The present study indicates that lexical access dynamics are largely affected by the perceptual clarity of the input, and not by the expectations. If this is so, those with assistive listening devices (e.g. CI) who must face consistently degraded input may automatically adjust their processing mechanisms to manage the degraded input (Farris-Trimble, McMurray and Rigler's, 2017). Therefore, understanding the underlying characteristics that manipulate the dynamic processing mechanisms is valuable both in the lab and the clinic.

Audiologists are the key mediators between those with hearing loss and listening in degraded speech conditions. How then do the current findings possibly alter their tactics in the clinic when providing listening techniques to their clients?

Due to the seemingly automatic and flexible nature of spoken word recognition in sentences, audiologists should explore assessing speech perception with more complex speech patterns found in sentences rather than simply words in isolation. In heightening the linguistic and perceptual complexity in the assessments, audiologists may have the opportunity to more concisely understand the challenges their clients are facing and what tactics are best to manage these situations.

Furthermore, CI users are often instructed in the clinic to more "effortfully listen" in noise. This tactic is often strenuous and tiring for the listener. In addition, the subsequent speech recognition from the listening strategy may not necessarily be more accurate, varying greatly between individuals and environments. Therefore, if these spoken word recognition mechanisms are more a natural consequence of degraded input and not an enforced strategy depending on training or hearing level, audiologists may be able to provide tactics that minimize effort and maximize listening capabilities.

Lastly, the automatic and flexible nature of spoken word recognition may bode well for those who are prelingually deafened and implanted at a later age. Those who are prelingually deafened and implanted later not only have reduced spoken language input early in life, but a decreased opportunity, when the brain is substantially more plastic, to effectively adjust listening strategies to fit the new input. Yet, spoken word recognition may be more flexible and allow for a wide-range of dynamic processes rather than a constrained few due to the nature of the individual or listening situation. This may allow those who are late implanted a better opportunity for speech perception than once thought. Further research needs to be done to understand the interactive nature of spoken word recognition in not only adults but children of varying hearing abilities, devices and conditions. However, this study has taken us a step closer in the direction or directions of understanding the complex nature of lexical access in degraded conditions.

Appendix A

3 Randomized Item-Sets (Target, Cohort, Unrelated, Unrelated)

	Target	Cohort	Unrelated	Unrelated
1	brick	bridge	toad	toast
2	peach	peak	rat	rag
3	coffee	coffin	tower	towel
4	chip	chin	dollar	dolphin
5	twig	twin	grape	grave
6	magnet	magic	pillow	pillar
7	page	paint	lettuce	letter
8	trash	trap	well	web
9	captive	captain	pencil	penny

10	rope	road	drug	drum
11	mustard	mustache	carrot	carriage
12	muffler	muffin	rabbit	racket
13	money	monkey	coast	coach
14	rocker	rocket	goal	goat
15	reach	read	cork	corn
1	magnet	magic	rocker	rocket
2	goal	goat	money	monkey
3	coffee	coffin	pillow	pillar
4	tower	towel	pencil	penny
5	twig	twin	carrot	carriage
6	grape	grave	rat	rag
7	page	paint	reach	read
8	trash	trap	dollar	dolphin
9	muffler	muffin	captive	captain
10	toad	toast	drug	drum
11	lettuce	letter	rope	road
12	well	web	cork	corn
13	rabbit	racket	mustard	mustache
14	chip	chin	peach	peak
15	brick	bridge	coast	coach
1	magnet	magic	lettuce	letter
2	pillow	pillar	captive	captain
3	coffee	coffin	goal	goat
4	tower	towel	pencil	penny
5	twig	twin	carrot	carriage
6	grape	grave	rat	rag
7	page	paint	reach	read
8	trash	trap	chip	chin
9	muffler	muffin	rocker	rocket
10	toad	toast	drug	drum

11	money	monkey	rabbit	racket
12	well	web	coast	coach
13	rope	road	peach	peak
14	dollar	dolphin	cork	corn

Appendix B

Carrier Sentences

1 on this screen please choose the

2 on this screen please select the

3 on this screen please find the

4 choose the picture of the

5 choose the image of the

6 for this set of items please find

7 for this set of items please choose

8 for this set of items please select

9 click on the picture of the

10 please select the image of the

11 please click on the image of the

12 please find the item

13 choose the image that is

14 please select the image closest to

15 find the picture of the

16 find the image of the

Appendix C

Images Used in VWP Task

Well
Chip
Captive
Drug
Chin
Captain
Drum
Twin

Twig
Goal
Pencil
Cork
Tower
Coffee
Pillow
Rabbit
Money
Rocker
Web

Goat
Penny
Corn
Towel
Coffin
Pillar
Racket
Monkey
Rocket
Rag

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*(4), 419-439.

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. *Journal of Speech Language and Hearing Research, 54*(1), 243. doi:10.1044/1092-4388(2010/09-0233)

Brouwer, S., & Bradlow, A. R. (2015). The temporal dynamics of spoken word recognition in adverse listening conditions. *Journal of Psycholinguistic Research, 45*(5), 1151-1160. doi:10.1007/s10936-015-9396-9

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and*

Cognitive Processes, 16, 507–534. http://dx.doi.org/10.1080/01690960143000074

Farris-Trimble A, McMurray B. The reliability of eye tracking in the Visual World Paradigm for the study of individual differences in real-time spoken word recognition. Journal of Speech Language and Hearing Research. in press.

Supplemental Material for The Process of Spoken Word Recognition in the Face of Signal Degradation. (2013). *Journal of Experimental Psychology: Human Perception and Performance*. doi:10.1037/a0034353.supp

Frauenfelder, U., Scholten, M., & Content, A. (2001). Bottom-up inhibition in lexical selection: Phonological mismatch effects in spoken word recognition. Language and Cognitive Processes, 16, 583–607. http://dx .doi.org/10.1080/01690960143000146

Fu, Q. J., & Nogaki, G., Quian-Jie. (2004) Noise susceptibility of cochlear implant users: The role of spectral resolution and smearing. *Journal of the Association for Research in Otolaryngology, 6*(19-27) DOI: 10.1007/s10162-004-5024-3

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word- recognition. Cognition, 25, 71–102. http://dx.doi.org/10.1016/0010- 0277(87)90005-9

Marslen-Wilson, W. D., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. Journal of Experimental Psychology: Human Perception and Performance, 15, 576–585. http://dx.doi.org/10.1037/0096-1523.15.3.576

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. Cognitive Psychology, 18, 1–86. http://dx.doi.org/10.1016/ 0010-0285(86)90015-0

McQueen, J. M., & Huettig, F. (2012). Changing only the probability that spoken words will be distorted changes how they are recognized. The Journal of the Acoustical

Society of America, 131(1), 509–517. http://dx.doi.org/10.1121/1.3664087.

McMurray, B. Nonlinear curvefitting for Psycholinguistics. Available from: https://osf.io/4atgv/.

Muller-Deiler, J., Schmidt, B. J., Rudert, H., Effects of noise on speech discrimination in cochlear implant patients. Ann. Otol. Rhinol. Laryngol. 166:303Y306, 1995.

Nelson, P. B., Jin S. H., Carney, A. E., Nelson D. A., Understanding speech in modulated interference: cochlear implant users and normal-hearing listeners. *Journal of Acoustics Society of America.* 113:961Y968, 2003.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115*(2), 357-395. http://dx.doi.org/10.1037/0033-295X.115.2.357

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19, 1–36. http://dx .doi.org/10.1097/00003446-199802000-00001

Reinisch, E., Jesse, A., & Mcqueen, J. M. (2010). Early use of phonetic information in spoken word recognition: Lexical stress drives eye movements immediately. *Quarterly Journal*

=

of Experimental Psychology, 63(4), 772-783. doi:10.1080/17470210903104412

Rigler, H., Farris-Trimble, A., Greiner, L., Walker, J., Tomblin, J. B., & McMurray, B. (2015). The slow developmental timecourse of real-time spoken word recognition. *Developmental Psychology, 51*(12), 1690-1703. http://dx.doi.org/10.1037/dev0000044.

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. http://dx.doi.org/10.1126/science.7777863