Stream Segregation and Speech Recognition in Noise in Individuals With Cochlear Implants

NCT04854031

9/29/2020

Study Protocol

Objectives

Individuals with cochlear implants struggle to understand speech in the presence of competing sound sources, although there is a wide range of individual variability in speech recognition outcomes across individuals (e.g. Friesen et al. 2001; Stickney et al. 2004). This variability can be attributed to individual differences in the quality of the auditory input that the listener can hear ("bottom-up" factors) and the cognitive abilities of the listener which facilitate the extraction of meaningful information from that input ("top-down" factors).

In quiet conditions, these factors explain much of the individual differences between cochlear implant users. "Bottom up" constraints on auditory resolution, reflected in sensitivity to temporal (Fu 2002; Luo et al. 2008) and spectral (Litvak et al. 2007; Saoji et al. 2009) modulations in acoustic signals, account for about half of the variance in cochlear implant listeners' speech recognition accuracy in quiet (Won et al. 2011). "Top down" cognitive measures believed to reflect individual differences in attention and fluid intelligence have also been linked to cochlear implant users' sentence recognition in quiet (e.g. Moberly and Reed 2019).

Interpreting the meaning of correlations between speech recognition and tests of cognitive function is challenging because individual differences in most cognitive tasks have a positive association with one another (Diamond 2013; Miyake and Friedman 2012; Shipstead et al. 2014) and with speech recognition (Dryden et al. 2017). The challenge in interpretation is that previously observed correlations may not arise from a direct relationship between a measured cognitive skill and speech recognition, but rather from an incidental link via some intermediate cognitive skill which is related to both the measured skill and speech recognition. Thus, it is possible that the use of isolated memory tasks (as in previous studies) fails to measure the critical individual differences that govern speech recognition in quiet and in noise. To address this limitation of previous work, this proposal includes a battery of working memory tasks designed to assess individual differences in the ability to store and process information in memory while minimizing confounds that arise from task-specific individual differences.

This proposal focuses on assessing the role of individual differences in working memory in speech recognition in noise. Working memory tasks have been previously shown to predict speech recognition in noise in adults with cochlear implants (Kaandorp et al. 2017; O'Neill, Kreft, and Oxenham 2019), although this relationship has not been consistently observed (Moberly, Houston, et al. 2017). Working memory tasks generally have high reliability (Wilhelm, Hildebrandt, and Oberauer 2013) and avoid issues with interpreting reaction time data (White, Curl, and Sloane 2016; White, Servant, and Logan 2018) and concerns about the validity of underlying constructs (Rey-Mermet, Gade, and Oberauer 2018) that are associated with tasks believed to measure attention.

Working memory is the ability to store and process sequences of information (Akeroyd 2008; Rönnberg et al. 2013). Storage and processing are distinct latent constructs (Unsworth et al. 2009), which are assessed to varying degrees by different memory tasks (Unsworth and Engle 2007). For example, reading span requires switching between processing sentences and recalling unrelated sequences of words or letters (Daneman and Carpenter 1980; Kane et al. 2004), and therefore primarily assesses processing. In contrast, digit span requires maintaining a sequence of numbers in memory and repeating them back in order, which primarily assesses storage. In this proposal we use a battery of quick, reliable working memory tasks that span the range from primarily assessing processing to primarily assessing storage. This will address a gap in current understanding of which elements of working memory facilitate reconstructing degraded speech.

This proposal extends the use of auditory and cognitive predictors to examining individual differences in the ability to track and understanding a target speaker in competing noise. In addition to the direct effects of auditory resolution and working memory on recognition of target speech, we propose that auditory resolution and cognitive factors also indirectly affect speech recognition in competing speech by determining the listener's ability to segregate target speech from competing sound sources. Competing speech can act in two ways: it can overlap with target speech and mask portions of the target which provide speech information ("energetic" masking), and it can intrude on the listener's processing of the target speech information ("informational" masking) (Kidd et al. 2008; Shinn-Cunningham and Best 2008). Individuals with normal hearing can rely on auditory cues, such as voice pitch, to help separate target and masking speech from one another, but individuals with cochlear implants have difficulty hearing voice pitch differences (Chatterjee and Peng 2008; Deroche et al. 2014) and need large differences between target and masker voices to recognize speech in noise (Meister et al. 2020).

Hypothesis: Stream segregation mediates the relationship between speech recognition in competing speech, auditory resolution and working memory

The premise of this proposal is that individual differences in auditory resolution and cognition affect speech recognition in competing speech directly by affecting recognition of target speech and indirectly by affecting the difficulty of segregating target speech from competing speech. These direct and indirect pathways can quantified in a partial mediator model (Baron and Kenny 1986), with stream segregation as the variable that mediates the predictive power of auditory resolution and cognition for speech recognition in competing speech. In this project, we will test whether a partial mediator model describes the relationship between speech recognition in competing speech, stream segregation, and auditory resolution and cognition.

Design

Sample Size and Power Analysis

Our preliminary data described below indicate that the measures of auditory resolution and working memory we use will correlate with sentence recognition accuracy with coefficients of at least r = 0.5. Replicating correlations of this magnitude in the proposed study with $\alpha = 0.05$ and $\beta = 0.80$ would require 29 participants. We do not currently have an estimate of the indirect path effect sizes in the mediator model because the stream segregation task we are proposing is novel. For a mediator analysis, the needed sample size to achieve $\beta = 0.80$ is at least 34 participants if the relationship between auditory resolution, working memory and stream segregation and the relationship between stream segregation and speech recognition in noise have effect sizes of 0.59 (Fritz and Mackinnon 2010). Larger sample sizes would be needed for smaller effect sizes. The limited timeline for this project makes it infeasible to power the study for smaller indirect path sizes (e.g. 71 participants if both effect sizes are 0.39), so the goal of this pilot project will be to provide an estimate of the magnitude of these effect sizes as a basis for determining feasibility of subsequent work. The mediator analysis will still be conducted, but interpretation of non-significant outcomes will be framed within the statistical power of the available sample.

Participants

30 participants who lost their hearing and received cochlear implants as adults will participate in this study. We will include unilaterally and bilaterally implanted individuals listening with their everyday hearing configuration. Individuals with residual acoustic hearing better than 60 dBA at any audiometric frequency will be excluded to avoid the confounding the study with the effects of bimodal hearing. Participants with residual hearing in the non-implanted ear will inset an earplug into that ear to further ensure the stimuli are inaudible to that ear. The upper age limit on participants will be 80 years of age, to avoid the rising prevalence of dementia at older ages (Piccirillo et al. 2008) while still including a wide age range. We will attempt to recruit equal numbers of men and women, although we will prioritize including all interested participants over achieving parity. All statistical analyses will be conducted with sex as an additional predictor, although we do not expect sex to affect experimental outcomes.

Methods

General methodology for at-home testing

To avoid the risk of transmitting COVID through the lab, experiments will be conducted by participants in their own homes while on a video call with lab staff. At-home testing will enable us to recruit more participants than just those that can commute to the lab. At least 10 participants in this study will be individuals who have previously completed some of these tasks in the lab, to enable comparison of in-lab and at-home performance.

Experiment setup and performance will be conducted through a HIPAA-compliant WebEx video call between the research participant and lab staff. Consent will be administered through the video call using forms available through the Boy Town National Research Hospital REDCap server. If participants have difficulty understanding speech through the video call the text chat feature will be used to communicate to the participant.

Prior to each experimental session, a loudspeaker, external sound card, and sound level meter will be shipped to the participant with set-up instructions. We anticipate that participants will need some assistance setting up equipment, which we provide time for after administering consenting. Once the experimenter verifies that the equipment is properly set up, we will direct them to a YouTube video which plays a speech-shaped noise that has the long-term average spectrum of the speech material and ask them to adjust the volume of their computer until the sound level meter reads 65 dB SPL (slow, A-weighted) at approximate ear-level, following instructions sent with the equipment.

Experiments will be hosted on the website cognition.run. This website hosts behavioral experiments written in JavaScript using the jsPsych library (de Leeuw 2015), which eliminates the need to download or install experimental software. Pre-recorded auditory stimuli are loaded at the start of each task to avoid interruptions due to buffering during playback. Auditory stimuli less than a few seconds in length can be dynamically generated at the start of each trial, which enables adaptive psychophysical testing. Visual stimuli are precisely timed, with stimulus transitions occurring within a few milliseconds of the intended transition time. Button and keyboard responses are recorded without any personally identifying information in a database on the website, and verbal responses will be recorded through the WebEx video call and stored on a secure database at Boys Town.

Primary Outcome: Speech recognition in two-talker masking speech

Perceptually Robust English Sentence Test Open-set (PRESTO) sentences (Tamati, Gilbert, and Pisoni 2013) will be presented in quiet and with a two-talker masker at an average target-to-masker ratio of +5 dB. Participants will hear lists 7 and 15 in quiet and 13 and 17 in noise, to limit the impact of cross-list differences in difficulty (Faulkner et al. 2015) on speech recognition accuracy. Performance will be quantified as the percentage of keywords correctly identified across sentences (152 keywords per condition).

For the two-talker masker, PRESTO sentences from unused sentence lists will be concatenated into a running speech stream. Two samples of this running stream will be sampled and superimposed to form the competing speech for each target sentence. Two talkers will be used to maximize the impact of informational masking on speech recognition (Freyman, Balakrishnan, and Helfer 2004). The masking speech will start 0.5 s prior to the target speech and stop 0.5 s after to ensure that participants are given a timing cue to distinguish the masker from the target. The combination of target sentence and maskers will be fixed ("frozen") across participants to account for fluctuations in the amount of energetic masking that arise from interactions between targets and maskers (Buss et al. 2020). We will fix long-term target-to-masker ratio because that is conventional in speech recognition research. In addition, we will calculate short-term envelope ratio, defined as the proportion of time in which the target speech has a larger instantaneous envelope than the masker for each sentence.

Predictor 1: Auditory resolution

Preliminary data from 20 participants with cochlear implants tested in the lab indicates that spectral and temporal modulation detection thresholds predict PRESTO sentence recognition accuracy in quiet. In a generalized linear regression model, spectral and temporal modulation detection thresholds accounted for $R^2 = 0.64$ of the variance in recognition accuracy (spectral $\beta = -0.76$, temporal $\beta = -0.47$). We will repeat these tasks in this study.

Modulation detection thresholds will be measured with a three-interval, three-alternative forced-choice task. The carrier signal is a 60 dB SPL, 400 ms long noise burst with energy between 350 and 6500 Hz (selected to match the stimuli in Litvak et al. 2007), gated on and off with 20 ms $\sin^2$ ramps. Carriers are modulated in the target interval and was unmodulated in the other two intervals. Feedback is given after each trial. In the temporal modulation detection task, the target interval is sinusoidally amplitude-modulated at a 100 Hz rate, which has been previously shown to correlate with speech recognition in listeners with cochlear implants when presented via direct stimulation (Fu 2002; Luo et al. 2008) and via a loudspeaker (Won et al. 2011). In the spectral modulation detection task, the target interval is spectrally modulated at 0.5 ripples per octave, which has been previously shown to correlate with speech recognition when presented via a loudspeaker (Anderson et al. 2012; Litvak et al. 2007; Saoji et al. 2009). Carrier level for every interval is randomly varied between -3 and +3 dB Sound Pressure Level (SPL) to diminish the availability of loudness differences between modulated and unmodulated intervals (McKay and Henshall 2010).

A 2-down, 1-up adaptive procedure will be used to measure modulation depth threshold, which converges on 71% detection accuracy (Levitt 1971). The temporal modulation detection task starts at a modulation depth of 100% and decreases in steps of 4 dB from the first to fourth reversal and 1 dB for the remaining reversals. The spectral modulation detection task starts at a peak-to-valley ratio of 40 dB and decreases in steps of 4 dB from the first to fourth reversal and 1 dB for the remaining reversals. Runs will end after 55 trials or 8 reversals. For each run, the final 4 reversals will be averaged to obtain the modulation threshold. Participants will complete three runs of each task. Outcomes will be the mean threshold across all runs of each task.

Predictor 2: Working memory

Preliminary data in young adults with normal hearing (Bosen and Doria, in review) demonstrate a link between a three different memory tasks and recognition of vocoded speech in the presence of a speech masker. These results indicate that the ability of young adults with normal hearing to recognize vocoded speech in speech is sensitive to individual differences in storage and processing. Digit span is also associated with vocoded sentence recognition in quiet in individuals with normal hearing (Bosen and Barry 2020).

In contrast, we did not find a relationship between PRESTO sentence recognition accuracy in quiet and digit span performance in 20 listeners with cochlear implants after controlling for auditory resolution ($r = 0.25$, $p = 0.28$). This finding differs from results in young adults with normal hearing listening to vocoded speech, but additional working memory tasks are needed to determine the reason for this difference. In this study, we will use a battery of quick, reliable memory tasks which assess both storage and processing to determine their relative importance when listening in quiet and in competing speech. These tasks include the three tasks shown to associated with vocoded speech recognition in competing speech:

Reading Span (primarily processing) – Participants alternate between judging if sentences are semantically sensible and storing letters in memory. Between 3 and 7 sentence-letter pairs are presented in each trial, after which participants recall the letters in order. The outcome is the total percentage of letters recalled in the correct position (Conway et al. 2005).

Serial Digit Span (primarily storage) – Participants hear a list of between 2 and 9 spoken digits presented with one second between stimulus onsets, then verbally repeat them in order (Bosen and Barry 2020; Bosen and Luckasen 2019). The outcome is the total percentage of digits recalled in the correct position.

Free Recall (primarily storage) – Participants read a list of 12 monosyllabic words shown for 750 ms with 250 ms between words, then repeat back words as they remember in any order. The outcome is the average number of words recalled from the list.

We will include a further two tasks to balance the distribution of tasks which primarily load onto storage and processing and to promote segregation of factors, because adding more measures facilitates latent factor identification (de Winter, Dodou, and Wieringa 2009). Digit updating is included to provide a task which primarily loads onto processing but does not require reading comprehension as reading span does. Running digit span requires both storage and processing (Shipstead et al. 2014) and also avoids linguistic variability.

Digit Updating (primarily processing) – Participants see between two and six boxes on screen. Sequences of digits are shown, one digit at a time, in a pseudorandom order across the boxes. Participants remember the most recent number shown in each box, and when the sequence ends, they must type in the last number that was shown in each box. The outcome is the percentage of last numbers correctly recalled across boxes. This task will be implemented as described in (Wilhelm, Hildebrandt, and Oberauer 2013).

Running Digit Span (storage and processing) – Between 12 and 20 spoken digits are presented at a rate of 4 per second, and participants recall as many digits as they can in order from the end of the list. The outcome is the average number of digits recalled in the correct position relative to the last item in the sequence across lists. This task will be implemented as described in (Cowan et al. 2005).

Predictor 3: stream segregation

A sequence of 6 spoken digits alternating in voice pitch (F0) between a high pitch and a low pitch will be presented. Participants will only repeat digits presented with a high pitch, requiring segregation of the two F0 streams. This is analogous to previous sequential streaming studies which used phonemes (David et al. 2017; Gaudrain et al. 2008), but the use of digits makes identification of individual items easier and insensitive to auditory resolution (Bosen and Luckasen 2019; Moberly, Harris, et al. 2017). Recalling 3 digits from each sequence minimizes variability in recall accuracy which arises at longer list lengths (Unsworth and Engle 2006). The number "seven" will be excluded because it is the only disyllabic digit in English, which can affect recall accuracy (Baddeley, Thomson, and Buchanan 1975; Egner, Sütterlin, and Lugo 2016). The sequence will be presented at a rate of 0.5 s per digit onset. This rate of presentation limits the available strategies to facilitate recall, which avoids introducing variance associated with individual differences in memory strategy (Bailey, Dunlosky, and Kane 2008, 2011). Sequential presentation of digits across streams avoids energetic masking that would arise if streams were presented concurrently. Digits spoken by one talker will be manipulated in Praat to have different F0s, thus keeping vocal tract length constant. Digits will alternate between F0 values of 85/200, 120/200, and 150/200 Hz, with the expectation that recall accuracy will increase as the F0 difference increases because the high and low pitch streams will be more segregated. We will also include a baseline condition in which all digits will be presented with an F0 of 200 Hz and participants recall every other digit, starting with the first. This baseline condition is expected to be more difficult than conditions with alternating voice pitches because no segregation cue is available. The difference in recall accuracy between alternating F0 conditions and the baseline condition will be the metric of stream segregation.