Blinky Dot 3.1 Manuscript

To do	Notes
Design	X
Get stims	X
Program	X
Make OSF page and submit prereg	X
IRB	X
Protocol & set up participant spreadsheets	X
Pilot (in lab)	X
Check data output from pilot	X
Set up youcanbook.me and calendars	X
Write audiocoding protocol	X
Print IRB and protocol	X

Introduction

A substantial body of work has indicated that speech recognition is improved when listeners can see as well as hear the talker, also known as visual enhancement (Norman P. Erber, 1969; Sumby & Pollack, 1954; Van Engen, Phelps, Smiljanic, & Chandrasekaran, 2014). Talking faces provide robust information about place of articulation (Binnie, Montgomery, & Jackson, 1974), a feature that is easily lost in noise or reverberation (Grant & Walden, 1996). The benefits of seeing and hearing a talker relative to hearing alone are quite robust and have been demonstrated in normal hearing children (Kirk et al., 2007), young adults (Van Engen et al., 2014), and older adults (Tye-Murray, Sommers, & Spehar, 2007b; Tye-Murray, Sommers, Spehar, Myerson, & Hale, 2010), also in hearing-impaired children (N. P. Erber, 1972; Kirk et al., 2007) and adults (Grant, Walden, & Seitz, 1998; Kaiser, Kirk, Lachs, & Pisoni, 2003).

Although it is clear that visual input can facilitate speech recognition, measures of spoken word identification accuracy provide an incomplete picture of the difficulty of the speech task. That is, these measures do not fully convey the cognitive resources necessary to process speech. To understand how visual information affects of speech identification, it is necessary to examine the "listening effort" (LE) used in processing such speech. The concept of LE relies on the assumption that humans possess a limited pool of cognitive resources (Kahneman, 1973), so as more cognitive resources are shifted toward more difficult listening tasks, fewer resources remain to quickly and accurately complete simultaneous tasks (see Pichora-Fuller et al., 2016). Consequently, one major way to quantify LE is by measuring performance on a secondary task while participants listen to speech. Listeners may be able to maintain

high levels of word recognition accuracy when having a conversation in either a quiet room or in a busy restaurant with competing speech and loud background noise, but the cognitive and attentional demands of listening in these two settings are likely to be quite different. In fact, a growing body of evidence suggests that the listening effort a listener expends may vary independently of word recognition accuracy. For example, Mackersie and Cones found that while performance accuracy remained the same during an easy, moderate, and difficult speech task, EMG and skin conductance indicated an increase in listening effort as task difficulty increased (Mackersie & Cones, 2011). Similarly, noise-reduction algorithms in hearing aids hearing may decrease LE without improving word recognition accuracy (Sarampalis, Kalluri, Edwards, & Hafter, 2009). Thus, while it is well-established that seeing a talker improves accuracy, these findings do not necessarily imply that the visual signal makes the task less cognitively effortful.

Previous findings on how seeing a talking face affects listening effort are mixed. Three studies, all of which used dual task paradigms, have found that participants expend more listening effort for audiovisual speech relative to audio-only speech (Fraser, Gagné, Alepins, & Dubois, 2010; Gosselin & Gagné, 2011). These findings of increased effort associated with seeing as well as hearing a talker may be a function of distraction, costs associated with simultaneously monitoring or processing two channels, or the act of integrating the stimuli. Studies using memory-based measures of listening effort have tended to show that seeing a talking face either reduces effort (Sommers & Phelps, 2016) or has no effect (Keidser, Best, Freeston, & Boyce, 2015).

Though this hypothesis was not rigorously tested, a pair of studies by Mishra and colleagues hint that the listening conditions may moderate the benefit of a visual signal (Mishra, Lunner, Stenfelt, Rönnberg, & Rudner, 2013a, 2013b). In one of their studies (Mishra et al., 2013a), they found that in speech-shaped background noise, scores on the CSCT were better in audiovisual conditions as opposed to audio-only conditions, but that in quiet conditions, scores on the CSCT were worse in audiovisual conditions, indicating that integrating visual and auditory signals incurs a cost, so it is not beneficial in quiet conditions.

However, in more challenging situations, the benefit of two modalities may outweigh the cost. In a recent study, Brown and Strand addressed this question using two different signal to noise ratios and two different LE paradigms: a recall task and a dual task (Violet A. Brown & Strand, 2019). While the recall paradigm was sensitive to changes in background noise, there was neither a significant main effect of modality nor a significant interaction between modality and SNR. In the dual task condition, however, a visual signal always increased listening effort, but this effect was less pronounced in the difficult signal to noise ratio. The pattern of data observed in the dual task experiment is similar to the one observed by Mishra et al. (Mishra et al., 2013b), and together these studies provide evidence that the cognitive costs of processing a visual signal may depend on the listening conditions. The discrepancy in the above findings may be explained, in part, by the considerable methodological variation present. For example, Mishra et al. (Mishra et al., 2013a) did not observe a reduction of LE in informational masking, despite that this effect was present in energetic masking. Similarly, Strand et al. observed entirely different patterns of data in the dual task and recall experiments, echoing previous evidence that LE measures do not necessarily measure the same underlying construct (Strand, Brown, Merchant, Brown, & Smith, 2018). Other research out of our lab (Violet A. Brown & Strand, 2019) found that in dual task paradigms, the cost of audiovisual processing was more pronounced in easy listening conditions, but this effect did not replicate in recall paradigms. These results suggest that methodological decisions matter; therefore, the variation in previous experiments poses a major challenge in reconciling their findings.

Also of interest to researchers is the benefit of other visual signals. Although most audiovisual speech perception research uses a clear, talking face as the visual signal, recognition benefits can also come from other types of visual input. For example, point-light displays (which provide crude information about articulation) can improve recognition (Rosenblum, Johnson, & Saldaña, 1996), as can faces with high levels of Gaussian blur (Tye-Murray, Spehar, Myerson, Hale, & Sommers, 2016). Both of these signals provide rough phonetic information in addition to accurate temporal information. Our lab recently conducted an experiment to assess the influence of non-phonetic features of the visual signal on LE.

In this study, the visual signal consisted of a modulating circle giving only information about the timing and amplitude of the acoustic speech (in other words, it cannot be lipread). The signal did not improve speech recognition or reduce listening effort in this study, contrary to hypotheses (Strand, Brown, & Barbour, 2018). In the second of two experiments, participants listened to individual words in two-talker babble while either viewing a stationary fixation dot (audio-only condition) or a dot modulating with the speech stream (signal condition). At the same time, participants repeated the words aloud and completed a semantic dual task (SDT; (Picou & Ricketts, 2014)) in which they determined whether the presented word was a noun as quickly as possible. While recognition accuracy did not differ between conditions, reaction times to complete the secondary task were significantly slower in the signal condition than in the audio-only condition, suggesting that the modulating circle increased the listening effort required to understand the speech stimuli. In a follow up study with older adults, the modulating circle still increased listening effort, but also increased spoken word identification accuracy (V. A. Brown, Strand, & Van Engen, 2019). The increase in spoken word recognition here may be due to the more difficult listening condition: in other words, the dot was providing superfluous information in easier listening conditions, so the benefit of the additional temporal information was not needed for speech recognition (Gosselin & Gagné, 2011).

To summarize, it is currently unclear how seeing a talking face affects listening effort relative to hearing alone, and the variety of methodological decisions in previous experiments makes it difficult to reconcile their results. An initial report (V. A. Brown et al., 2019) indicates that visual cues to temporal information may reduce LE in difficult listening situations, but what remains unclear is how the effect of a talking face compares with other visual signals. Given the ubiquity of audiovisual speech and the fact that maintaining high levels of listening effort can lead to mental fatigue or distress, understanding how visual information affects listening effort has important practical and clinical consequences. In the present study, we aim to conduct the most comprehensive evaluation to date of how visual information affects LE by manipulating visual stimulus type, task difficulty, and type of masking noise.

The current study presented two types of visual stimuli in addition to an audio only (AO) condition: a dot that modulated in size that corresponds to the amplitude of the speech stimulus (“signal condition,” similar to the stimuli used in (Strand, Brown, & Barbour, 2018)), and a video of the head and shoulders of a talking face (“audiovisual condition”). The talking face and modulating dot provide different types of information about the auditory stimulus. The dot is simple and provides low-level, temporal input, whereas the face is much more visually complex and conveys higher-level phonetic information, as well as information about timing. Thus, the two types of visual stimuli may be expected to provide different amounts of information about speech and incur different LE costs. Including both types of stimuli in the same study will shed light on the features of visual input that contribute to changes in both recognition accuracy and LE.

Given our previous findings (Brown & Strand, 2019) and previous research (Gosselin & Gagné, 2011), we also expect that the extent to which the visual stimuli will affect LE will vary as a function of difficulty. When speech recognition performance is already at a high level, visual information is not necessary for recognition, so the signal may actually incur a cost to process (Gosselin & Gagné, 2011), and consistent with (Strand, Brown, & Barbour, 2018)). An interaction with difficulty has been suggested by prior work that showed that participants had fewer available cognitive resources when the speech was accompanied by visual information (similar to the findings of (Brown & Strand, 2019; Gosselin & Gagné, 2011) in easy conditions, but in more difficult listening conditions participants had more spare cognitive resources, or a decrease in LE (Mishra et al., 2013a). The contradictory findings in the literature about whether visual information increases or decreases LE may therefore be attributed the difficulty of the listening task. Thus, we will also present the study in different signal to noise ratios (SNRs) in order to manipulate the difficulty of the listening task.

Participants will be randomly assigned to complete Experiment 1 or Experiment 2, which use informational masking (IM; two-talker babble) and energetic masking (EM; speech-shaped noise), respectively. In informational masking, both the target stream and masking stream are well represented in the auditory system, but intelligibility is impaired by the difficulty of segregating target stream from the babble streams (e.g., (Freyman, Balakrishnan, & Helfer, 2004; Freyman, Helfer, McCall, & Clifton, 1999)). Energetic masking impairs the intelligibility of the speech by overlapping with pieces of target speech and reducing peripheral audibility. Previous studies have shown that the underlying mechanisms of informational and energetic masking are partially differentiable. Van Engen and colleagues (Van Engen, Chandrasekaran, & Smiljanic, 2012) found that speech recognition accuracy in energetic masking conditions did not predict recognition accuracy in informational masking conditions. Additionally, previous work has found that better working memory predicted speech recognition in informational masking conditions, but not energetic masking conditions (Koelewijn, Zekveld, Festen, Rönnberg, & Kramer, 2012; Zekveld, Festen, & Kramer, 2013). Thus, informational maskers place greater demands on executive functions and interfere more substantially with central processing compared to energetic maskers.

We can use these previous findings to generate hypotheses about how a visual signal affects listening effort in each condition. Given the robust literature showing that background noise affects listening effort (Strand, Brown, Merchant, et al., 2018), we expect to see an effect of SNR on participant reaction time to the secondary task in both energetic and informational masking -- harder SNRs will lead to slower reaction times on the secondary task. We also expect that there will be lower accuracy in harder SNRs in both masking types.

We hypothesize that in informational masking, there will be a significant effect of condition on reaction time and accuracy: there will be slower responses for the signal as compared to the A-only condition, and for the AV condition as compared to the signal condition (consistent with Brown et al., 2019; Strand, Brown, & Barbour, 2018). However, we also expect to see a stepwise increase in accuracy -- A-only will have the worst accuracy, followed by the signal condition, and then the audiovisual condition.

In terms of the interaction between SNR and condition in informational masking, we expect that the visual conditions (signal and AV) will hurt more in the easier listening condition, because they don't provide as much meaningful information. We expect that this interaction will result in larger differences in RTs between the visual conditions and A-only in the easy SNR, and larger differences in word

identification between the visual conditions and A-only in the hard SNR (the visual conditions are more beneficial in harder conditions because they provide complementary information).

We hypothesize that in energetic masking, there will also be a significant effect of condition on reaction time: just as in informational masking, there will be the slowest responses for AV, followed by signal, and fastest for a-only. However, we predict that there will be little or no recognition benefit for the dot in energetic masking, because the dot does not provide any information that is being masked. Therefore, we expect the A-only and signal conditions to have similar word recognition accuracy, while the AV condition will have better recognition.

In terms of the interaction between SNR and condition in energetic masking, we expect that with reaction times, the difference between A-only and the two visual conditions will be larger in the easy SNR. We expect that there will be a larger difference between A-only and AV in terms of recognition accuracy in the hard SNR, but we do not expect the dot to be helpful in the difficult listening condition.

Our hypotheses indicate that masking conditions may further moderate the interactive effect of background noise and a visual signal on listening effort. Therefore, we also plan to test the three-way interaction between visual condition, SNR, and masker type. We hypothesize that this interaction will be significant; however, of utmost interest in the current study are the planned comparisons outlined above.

Experiment 1 Method

Participants

A total of XX native English speakers, ages 18–23 years, with self-reported normal hearing and normal or corrected-to-normal vision were recruited from the Carleton College community. Participants provided written consent and received $11 for 60 minutes of participation. After following the exclusion criteria outlined below, only XX of the participants' data were analyzed. Carleton College's Institutional Review Board approved all research procedures.

Stimuli

Both experiments use the semantic dual-task (SDT) to assess listening effort (Picou & Ricketts, 2014; Strand, Brown, & Barbour, 2018; Strand, Brown, Merchant, et al., 2018). In this task, participants listen to and repeat aloud individual words in a continuous stream, and are asked to press a button on a Cedrus RB-740 buttonbox as quickly and accurately as possible if the word can be classified as a noun. Speech stimuli consisted of 600 words selected from a subset of the SUBTLEX-US database (Brysbaert, New, & Keuleers, 2012 see Strand et al., 2018 for more details on stimulus selection). To remain consistent with previous experimentation (Picou & Ricketts, 2014; Strand, Brown, & Barbour, 2018), 55% of words selected were classified as nouns according to SUBTLEX-US part of speech dominance data (Brysbaert et al., 2012). The words selected were two syllables or less, included two to five phonemes, had 3 or greater log frequences, and did not include articles, conjunctions, or names.

A female native English speaker without a discernible regional accent produced all target words. Auditory stimuli were recorded at 16-bit, 44100 Hz using a Shure KSM-32 microphone with a plosive screen, and were edited and equated for root-mean-square amplitude using Adobe Audition prior to being combined with the corresponding visual signal. Visual stimuli were recorded with a Panasonic AG-AC90 camera. Videos were edited with iMovie (version 10.1). Two-talker babble was presented continuously as an informational masker for the target words. The target speech was delivered binaurally at approximately XX dB SPL and noise at XX dB SPL (SNR = -6 dB, "hard" condition) or at XX dB SPL (SNR = 14 dB,

"easy" condition) via Sennheiser HD 280 Pro headphones. Stimuli were presented on a 21.5-inch iMac computer via SuperLab 6 (Cedrus).

We used a custom Java program to produce the circle which appears at target onset, modulated between 50 and 200 pixels (approximately 1.1–4.5 cm) with the amplitude of speech (along with a luminance change varying from 39% software luminance corresponding to 0% sound level to 100% software luminance corresponding to 100% sound level), and disappeared at offset. Videos of the Java program were collected using QuickTime screen capture and presented in SuperLab so that reaction time data could be collected.

The 600 words were divided into 6 lists balanced to maintain a 55% noun composition. Two lists, one for each SNR, were made for each of the three visual conditions: *audio-only*, *signal*, and *face*. In all conditions words were presented in continuous energetic masking. In the *audio-only* condition, the screen contained a static dot that remained unchanged between trials. In the *signal* condition, words were presented simultaneously with the modulating circle, and in the *face* condition, the words were presented with the accompanying video of the speaker saying the word. The word lists were counterbalanced across conditions such that over the course of the experiment, participants heard each word exactly once, and each list of words appeared equally often in each of the six conditions.

Design and Procedure

Participants sat a comfortable distance from the 21.5-inch iMac computer and completed 6 SDT blocks (SDT + *audio only,* SDT + *signal,* and SDT + *face*, each at SNR = -6 dB and SNR = 14 dB). The order of the blocks was counterbalanced across participants. During SDT blocks, participants were instructed to listen to a stream of words and press a button as quickly and accurately as possible if they determined the word was a noun. After making the noun judgement, participants repeated the word aloud regardless of its part of speech. Reaction times to trials in which the participants classified the word as a noun were used as a measure of LE. Accuracy for the noun classification was not scored (see Picou & Ricketts, 2014; Strand, Brown, & Barbour, 2018). In all trials, the interstimulus interval varied between 2,000 and 4,500ms in 250ms intervals. Accuracy at the speech identification task was scored offline by research assistants. After each block, the participants were presented with the NSAS-TLX (Hart & Staveland, 1988; Seeman & Sims, 2015). The NASA-TLX uses a 21-point scale to measure task difficulty. Participants were instructed to type in a number for 1-21 into a textbox to rate how difficult they perceived the previous block to be.

Exclusion criteria

Participants (N = X) were excluded from all analyses if, for any condition, their accuracy on the word recognition task is worse than three standard deviations below the mean accuracy for that condition. We also excluded participants (N = Y) from all analyses if their reaction time on noun classifications, for any condition, was more than three standard deviations above or below the mean for that condition. Finally, we replaced all participants who encountered technical difficulties such as computer crashes while testing (N = X).

Individual SDT trials were excluded if the participant's reaction time for that trial was outside of the response window (i.e. more than 4000 ms), or if the participant's reaction time was more than three median absolute deviations above or below that participant's median reaction time. Following the recommendations of (Leys, Ley, Klein, Bernard, & Licata, 2013), we use the absolute deviation around

the median rather than the standard deviation of the mean because individual reaction times tend to be skewed. For roughly XXX trials in the combined dataset, participants inadvertently pressed the button more than once. Although participants were not presented with the stimulus more than once, the data file shows an additional instance of that stimulus. Rather than defining an arbitrary criterion for which instances to remove, we chose to exclude these trials.

Results

All analyses were conducted using linear mixed effects modeling via the *lme4* package in R, version 3.5.1 (Bates et al., 2014). We used the maximal random effects structure justified by the design (Barr, Levy, Scheepers, & Tily, 2013), and we compared models using likelihood ratio tests. More information about random effects structures and issues of model convergence is available in the R script at https://osf.io/cyrhb/.

Hypothesis 1.

Experiment 2 Method

Participants

XX native English speakers, ages 18–23 years, with self-reported normal hearing and normal or corrected-to-normal vision were recruited from the Carleton College community. Participants provided written consent and received $10 for 60 minutes of participation. After the exclusion criteria, only XX of the participants' data were analyzed. Carleton College's Institutional Review Board approved all research procedures.

Stimuli

Stimuli were identical to those in Experiment 1, and were presented in the same manner at the same SNRs with the exception of the masking. Words were presented in energetic, speech-shaped noise that matched the long-term average spectrum of the speech files. The two levels of background noise match the two SNRs in Experiment 1.

Design and Procedure

The procedures in Experiment 2 were identical to those in Experiment 1.

Results

MANUSCRAPS

Ways in which BD3.0 was different:

- SNR = -7 dB and SNR = 9 dB
    - Speech: -27 RMS
    - Easy babble: -36
    - Hard babble: -20
- interstimulus interval varied between 2,000 ms, 2,500 ms, and 3,000 ms

BD3.1 stims

- Speech: in praat, 61.93
- Easy noise: in praat, 47.93
    - SNR: +14
- Hard noise: in praat, 67.93
    - SNR: -6

Notes on program:

- Piloted Nov 19, everything ran smoothly
- Tested Jan 10, crashed every time during AV files (and sometimes triggered kernel panic). Using SL6 and RB-740
- Jan 13 and 14:
    - Randy updated driver for button box
    - JS realized it was only crashing for face (.mp4) and not signal (.mov) files, so converted all .mp4 files to .mov and reduced file size (in case it is a memory issue)
    - L booth still crashes for only AV files (energetic group 3, on the third block which is AV easy)
    - JS remade most of the blocks in SL6 (didn't change anything deliberately, just started from one block that worked and copied it/reconnected) in informational, then copied and changed noise to make new energetic as well. Didn't fix anything
    - Maybe it's a driver problem? SL5 uses the VPC driver (which is built-in on apple) and SL6 uses the D2xx and they're mutually exclusive. Randy suppressed the VPC driver and it still doesn't work.
    - JS changed the program so use keyboard rather than button box. Still doesn't work (so must not be about the driver?)
    - Uninstalled everything about superlab 6 (including D2xx driver) from right booth, reprogrammed everything in superlab 5.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3). https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R., Singmann, H., … Green, P. (2014). *Package "lme4."* Retrieved from R foundation for statistical computing, Vienna, 12. website: https://github.com/lme4/lme4/

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of Speech and Hearing Research*, *17*(4), 619–630.

Brown, V. A., & Strand, J. F. (2019). About Face: Seeing the Talker Improves Spoken Word Recognition but Increases Listening Effort. *Journal of Cognition*, *2*(1). https://doi.org/10.5334/joc.89

Brown, V. A., Strand, J., & Van Engen, K. (2019). *A modulating circle reduces listening effort and improves spoken word recognition in older adults*. Retrieved from https://psyarxiv.com/v8x4q/download?format=pdf

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101–1109.

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, *110*(5 Pt 1), 2527–2538.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, *44*(4), 991–997.

Cooke, M., Garcia Lecumberri, M. L., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *The Journal of the*

Acoustical Society of America, *123*(1), 414–427.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, *12*(2), 423–425.

Erber, N. P. (1972). Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing. *Journal of Speech and Hearing Research*, *15*(2), 413–422.

Fraser, S., Gagné, J.-P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues. *Journal of Speech, Language, and Hearing Research: JSLHR*, *53*(1), 18–33.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2001). Spatial release from informational masking in speech recognition. *The Journal of the Acoustical Society of America*, *109*(5), 2112–2122.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *The Journal of the Acoustical Society of America*, *115*(5 Pt 1), 2246–2256.

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *The Journal of the Acoustical Society of America*, *106*(6), 3578–3588.

Gosselin, P. A., & Gagné, J.-P. (2011). Older adults expend more listening effort than young adults recognizing audiovisual speech in noise. *International Journal of Audiology*, *50*(11), 786–792.

Grant, K. W., & Walden, B. E. (1996). Evaluating the articulation index for auditory-visual consonant recognition. *The Journal of the Acoustical Society of America*, *100*(4), 2415–2424.

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America*, *103*(5 Pt 1), 2677–2690.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of

empirical and theoretical research. *Advances in Psychology*, *52*, 139–183.

Helfer, K. S., & Freyman, R. L. (2005). The role of visual speech cues in reducing energetic and informational masking. *The Journal of the Acoustical Society of America*, *117*(2), 842–849.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Kaiser, A. R., Kirk, K. I., Lachs, L., & Pisoni, D. B. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing Research: JSLHR*, *46*(2), 390–404.

Keidser, G., Best, V., Freeston, K., & Boyce, A. (2015). Cognitive spare capacity: evaluation data and its association with comprehension of dynamic conversations. *Frontiers in Psychology*, *6*, 597.

Kirk, K. I., Hay-McCutcheon, M. J., Holt, R. F., Gao, S., Qi, R., & Gehrlein, B. L. (2007). Audiovisual spoken word recognition by children with cochlear implants. *Audiological Medicine*, *5*(4), 250–261.

Koelewijn, T., Zekveld, A. A., Festen, J. M., Rönnberg, J., & Kramer, S. E. (2012). Processing load induced by informational masking is related to linguistic abilities. *International Journal of Otolaryngology*, *2012*, 865731.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766.

Mackersie, C. L., & Cones, H. (2011). Subjective and psychophysiological indexes of listening effort in a competing-talker task. *Journal of the American Academy of Audiology*, *22*(2), 113–122.

Mishra, S., Lunner, T., Stenfelt, S., Rönnberg, J., & Rudner, M. (2013a). Seeing the talker's face supports executive processing of speech in steady state noise. *Frontiers in Systems Neuroscience*, *7*, 96.

Mishra, S., Lunner, T., Stenfelt, S., Rönnberg, J., & Rudner, M. (2013b). Visual information can hinder working memory processing of speech. *Journal of Speech, Language, and Hearing Research*, *56*, 1120–1132.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., … Wingfield, A. (2016). Hearing impairment and cognitive energy: The Framework for Understanding Effortful Listening (FUEL). *Ear and Hearing*, *37 Suppl 1*, 5S – 27S.

Picou, E. M., & Ricketts, T. A. (2014). The effect of changing the secondary task in dual-task paradigms for measuring listening effort. *Ear and Hearing*, *35*(6), 611–622.

Rabbitt, P. M. (1968). Channel-capacity, intelligibility and immediate memory. *The Quarterly Journal of Experimental Psychology*, *20*(3), 241–248.

Rosenblum, L. D., Johnson, J. A., & Saldaña, H. M. (1996). Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech and Hearing Research*, *39*(6), 1159–1170.

Sarampalis, A., Kalluri, S., Edwards, B., & Hafter, E. (2009). Objective measures of listening effort: Effects of background noise and noise reduction. *Journal of Speech, Language, and Hearing Research: JSLHR*, *52*(5), 1230–1240.

Seeman, S., & Sims, R. (2015). Comparison of psychophysiological and dual-task measures of listening effort. *Journal of Speech, Language, and Hearing Research: JSLHR*, *58*(6), 1781–1792.

Sommers, M. S., & Phelps, D. (2016). Listening effort in younger and older adults: A comparison of auditory-only and auditory-visual presentations. *Ear and Hearing*, *37 Suppl 1*, 62S – 8S.

Strand, J. F., Brown, V. A., & Barbour, D. L. (2018). Talking points: A modulating circle reduces listening effort without improving speech recognition. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-018-1489-7

Strand, J. F., Brown, V. A., Merchant, M. B., Brown, H. E., & Smith, J. (2018). Measuring listening effort: Convergent validity, sensitivity, and links with cognitive and personality measures. *Journal of Speech, Language, and Hearing Research: JSLHR*, *61*, 1463–1486.

Sumby, W. H., & Pollack, I. (1954). Visual contributions to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215.

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007a). Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends in Amplification*, *11*(4), 233–241.

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007b). The effects of age and gender on lipreading abilities. *Journal of the American Academy of Audiology*, *18*(10), 883–892.

Tye-Murray, N., Sommers, M. S., Spehar, B., Myerson, J., & Hale, S. (2010). Aging, audiovisual integration, and the principle of inverse effectiveness. *Ear and Hearing*, *31*(5), 636–644.

Tye-Murray, N., Spehar, B., Myerson, J., Hale, S., & Sommers, M. S. (2016). Lipreading and audiovisual speech recognition across the adult lifespan: Implications for audiovisual integration. *Psychology and Aging*, *31*(4), 380–389.

Van Engen, K. J., Chandrasekaran, B., & Smiljanic, R. (2012). Effects of speech clarity on recognition memory for spoken sentences. *PloS One*, *7*(9), e43753.

Van Engen, K. J., Phelps, J. E. B., Smiljanic, R., & Chandrasekaran, B. (2014). Enhancing speech intelligibility: Interactions among context, modality, speech style, and masker. *Journal of Speech, Language, and Hearing Research: JSLHR*, *57*(5), 1908–1918.

Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2013). Task difficulty differentially affects two measures of processing load: the pupil response during sentence processing and delayed cued recall of the sentences. *Journal of Speech, Language, and Hearing Research: JSLHR*, *56*(4), 1156–1165.

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of speech and hearing research*, *17*(4), 619-630.

Erber, N. P. (1972). Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing. *Journal of Speech and Hearing Research*, *15*(2), 413-422.

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America*, *103*(5), 2677-2690.

Kaiser, A. R., Kirk, K. I., Lachs, L., & Pisoni, D. B. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing Research*.

Keidser, G., Best, V., Freeston, K., & Boyce, A. (2015). Cognitive spare capacity: evaluation data and its association with comprehension of dynamic conversations. *Frontiers in Psychology*, *6*, 597.

Kirk, K. I., Hay-McCutcheon, M. J., Holt, R. F., Gao, S., Qi, R., & Gerlain, B. L. (2007). Audiovisual spoken word recognition by children with cochlear implants. *Audiological Medicine*, *5*(4), 250-261.

Koelewijn, T., Zekveld, A. A., Festen, J. M., Rönnberg, J., & Kramer, S. E. (2012). Processing load induced by informational masking is related to linguistic abilities. *International journal of otolaryngology*, *2012*.

Mackersie, C. L., & Cones, H. (2011). Subjective and psychophysiological indexes of listening effort in a competing-talker task. Journal of the American Academy of Audiology, 22(2), 113-22.

Tye-Murray, N., Sommers, M. S., Spehar, B., Myerson, J., & Hale, S. (2010). Aging, audiovisual integration, and the principle of inverse effectiveness. *Ear and hearing*, *31*(5), 636.

Van Engen, K. J. (2012). Speech-in-speech recognition: A training study. *Language and Cognitive Processes*, *27*(7-8), 1089-1107.

Zekveld, A. A., Rudner, M., Johnsrude, I. S., & Rönnberg, J. (2013). The effects of working memory capacity and semantic cues on the intelligibility of speech in noise. *The Journal of the Acoustical Society of America*, *134*(3), 2225-2234.