Psychonomic Bulletin & Review (2019) 26:291–297
https://doi.org/10.3758/s13423-018-1489-7

BRIEF REPORT

Talking points: A modulating circle reduces listening effort without improving speech recognition

Julia F. Strand[1] · Violet A. Brown[1] · Dennis L. Barbour[2]

Published online: 22 May 2018

Abstract

Speech recognition is improved when the acoustic input is accompanied by visual cues provided by a talking face (Erber in *Journal of Speech and Hearing Research, 12*(2), 423–425 1969; Sumby & Pollack in *The Journal of the Acoustical Society of America, 26*(2), 212–215, 1954). One way that the visual signal facilitates speech recognition is by providing the listener with information about fine phonetic detail that complements information from the auditory signal. However, given that degraded face stimuli can still improve speech recognition accuracy (Munhall et al. in *Perception & Psychophysics, 66*(4), 574–583, 2004), and static or moving shapes can improve speech detection accuracy (Bernstein et al. in *Speech Communication, 44*(1/4), 5–18, 2004), aspects of the visual signal other than fine phonetic detail may also contribute to the perception of speech. In two experiments, we show that a modulating circle providing information about the onset, offset, and acoustic amplitude envelope of the speech does not improve recognition of spoken sentences (Experiment 1) or words (Experiment 2), but does reduce the effort necessary to recognize speech. These results suggest that although fine phonetic detail may be required for the visual signal to benefit speech recognition, low-level features of the visual signal may function to reduce the cognitive effort associated with processing speech.

Keywords Spoken word recognition · Speech perception · Cross-modal attention

Recognizing speech in noisy or degraded conditions is a difficult perceptual task that is facilitated when the acoustic input is accompanied by visual cues provided by the talking face. Numerous studies have demonstrated "visual enhancement" by showing that adult listeners correctly identify more words when they can see and hear the talker relative to hearing alone (Erber, 1969; Sumby & Pollack, 1954). Although this research highlights the benefit of audiovisual speech, it remains unclear precisely what information a talking face conveys. The visual signal certainly provides complementary phonetic information to the auditory signal, such as cues about place of articulation—a feature that is easily lost in noisy or reverberant conditions (Grant & Walden, 1996). However, visual input may also provide valuable information other than fine phonetic detail. For example, coarse visual signals that omit much of the detail of talking faces—including point-light displays (Rosenblum, Johnson, & Saldaña, 1996), faces viewed at large distances (Jordan & Sergeant, 2000), and faces viewed across a range of spatial frequencies (Munhall, Kroos, Jozan, & Vatikiotis-Bateson, 2004)—still result in visual enhancement. Thus, features of visual stimuli other than fine-grained cues to phonetic content may also facilitate speech recognition.

In addition to the research on speech *recognition*, some research suggests that visual signals also facilitate speech *detection*. In detection studies, listeners must simply determine whether or not speech is present in high levels of background noise, rather than identify the speech. Although research on recognition tends to focus on the role of fine phonetic detail in visual enhancement, detection research has emphasized the contribution of attentional and temporal components of the visual signal. For example, although a talking face is most successful at reducing the detection threshold, other types of visual speech stimuli can improve listeners' ability to detect speech in noise: a static rectangle that appears at the

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13423-018-1489-7) contains supplementary material, which is available to authorized users.

✉ Julia F. Strand
jstrand@carleton.edu

[1] Department of Psychology, Carleton College, Northfield, MN, USA

[2] Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, MO, USA

onset and disappears at the offset of the speech, a dynamic Lissajous figure (i.e., a dynamic horizontal oval) that grows and shrinks vertically with the amplitude of the acoustic signal, and a low-contrast face all reduce the detection threshold relative to the audio-only threshold (Bernstein, Auer, & Takayanagi, 2004; Tye-Murray, Spehar, Myerson, Sommers, & Hale, 2011).

These results suggest that abstract visual stimuli are sufficient to facilitate detection of speech in noise. Therefore, in addition to fine phonetic detail, the visual signal may provide the listener with temporal information indicating the onset and offset of the speech stream, and may also direct the listener's attention to salient moments in the auditory signal. Although this previous research demonstrates that a dynamic or static figure other than a mouth can enhance detection, it is unclear whether these nonface figures helped the listener *recognize* the content of the speech. Only two studies have tested whether temporal cues from abstract visual stimuli can facilitate recognition (Schwartz, Berthommier, & Savariaux, 2004; Summerfield, 1979), and both found no evidence of visual enhancement. However, audiovisual asynchrony of just 40 milliseconds (ms) has been shown to eliminate visual enhancement effects in detection studies (Kim & Davis, 2004), so any asynchrony, even that which is consciously undetectable (Grant, van Wassenhove, & Poeppel, 2004), may interfere with visual enhancement. Technological improvements since Summerfield (1979) may provide more precise temporal alignment between the auditory and visual signals, allowing visual enhancement effects to emerge. Further, if the benefits provided by abstract visual stimuli are relatively small, they may require a highly powered study in order to be detected, and both prior studies had small sample sizes ($N < 13$). Given the robust benefits of seeing a talking face on speech recognition and the fact that abstract visual stimuli can benefit speech detection, we hypothesized that an abstract, modulating visual stimulus that lacks phonetic detail but provides precise temporal cues about the acoustic signal would facilitate speech recognition.

Experiment 1

Method

All stimuli, raw data, code for analysis, and software for creating the visual stimuli are available online at https://osf.io/b94yx/.

Participants

One hundred sixty-six 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 $5 for 30 minutes of participation. Carleton College's Institutional Review Board approved all research procedures.

Stimuli

Stimuli were selected from the Speech Perception in Noise (SPIN) database (Kalikow, Stevens, & Elliott, 1977). We included both high-predictability (HP) and low-predictability (LP) sentences to assess whether any effect of the visual signal depends on predictability (see Van Engen, Phelps, Smiljanic, & Chandrasekaran, 2014 for evidence of greater visual enhancement from a face for semantically constrained sentences), and presented sentences in two-talker babble (see Helfer & Freyman, 2005 for evidence of greater visual enhancement in two-talker babble than steady-state noise). A female native English speaker without a strong regional accent produced all target sentences. Stimuli were recorded at 16-bit, 44100 Hz using a Shure KSM-32 microphone with a plosive screen, and were edited and equated for RMS amplitude using Adobe Audition prior to being combined with the corresponding visual signal. The target speech was delivered binaurally at approximately 66 dB SPL and noise at 70 dB SPL (SNR = −4 dB) via Sennheiser HD 280 Pro headphones. We used a custom JavaScript program to create four types of visual stimuli: *audio-only*, *static*, *signal*, and *yoked* (see Table 1 for descriptions, and Supplementary Materials for examples of each type).

In all conditions, the visual stimulus appeared as a small, filled-in circle. In the conditions in which the circle was modulated (*signal* and *yoked*), the diameter ranged from 50 to 200 pixels (approximately 1.1–4.5 cm), the amount of time between graphics updates (i.e., the time step) was 50 ms, and the average size of the moving low-pass filter for the acoustic signal was 151 samples. In the conditions in which the circle was unmodulated (*audio-only* and *static*), the diameter was fixed at 50 pixels. When the circle diameter was modulated, the luminance of the circle also changed linearly as a function of the amplitude of the acoustic signal, with 100% software luminance corresponding to 100% software sound level and 39% software luminance corresponding to 0% software sound level (i.e., silence). When unmodulated, the circle remained at 39% software luminance. The luminance manipulation was included to more effectively draw the listener's attention to salient moments in the auditory stream.

Design and procedure

Each participant was randomly assigned to one of the four conditions. Participants sat a comfortable distance from a 21.5-inch iMac computer, and were presented with the same 140 target sentences in a pseudorandomized order (70 HP and

Table 1 Four conditions of Experiment 1

Condition	Description	Visual information provided
Audio only	Circle remained on and unmodulated throughout the entire experiment	Nothing
Static	Circle appeared at target onset, remained unmodulated, and disappeared at target offset.	Target onset and offset
Signal	Circle appeared at target onset, grew and shrank with the amplitude of the acoustic envelope of the target speech stream, and disappeared at target offset	Target onset, modulation, and offset
Yoked	Circle appeared at target onset and was modulated based on a sentence other than the target sentence	Target onset; included to determine whether the listener was extracting meaningful information from the visual signal or simply attending more closely to the acoustic signal in the presence of a dynamic figure

70 LP, intermixed) in a continuous stream of two-talker babble. Participants were asked to type the target sentence in a response box and then press enter, and were encouraged to guess when unsure. Participants were instructed to continue looking at the screen throughout the experiment because the circle may provide helpful cues about the contents of the target speech. The onset of the speech began a variable amount of time (1,500 ms–3,000 ms in 500 ms steps) after the end of the previous trial.

Responses were scored off-line by research assistants. We analyzed recognition accuracy for both the full sentences (given that information about speech onset is likely to be most helpful for items early in the sentence) and sentence-final words (to assess whether the visual signal benefits HP words more than LP words; see Van Engen et al., 2014). The first three sentences of the pseudorandomized list were counted as practice, and were therefore not included in the analyses. At the end of the study, participants were asked to subjectively rate task difficulty and perceived accuracy at completing the task. Because this was not of primary importance to the experiment, these results can be found in the Supplementary Materials.

Results and discussion

Responses were corrected for obvious typographical and spelling errors, and homophones were counted as correct. Responses that contained both words of a contraction (e.g., "I have") were scored as correct for the single contracted word. Articles ("the," "a," "an") were excluded from analysis, and compound words (e.g., "bullet-proof," "household," "policeman") were coded as two separate words. One participant was excluded from all analyses due to low accuracy (less than three *SD*s below the mean), so the final analysis included 165 participants.

Data were analyzed using linear mixed-effects models via the lme4 package in R (Version 3.3.3; Bates et al., 2014). To determine whether condition affected accuracy, we first built

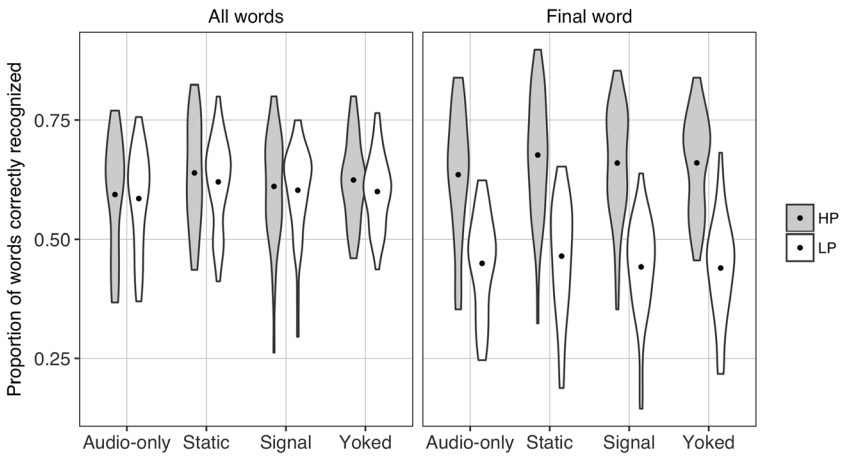

Fig. 1 Violin plots showing the distribution of participant mean accuracies by condition and type for the analysis of all words (left) and sentence-final words (right). The dot shows the mean value in each condition, and the width depicts the density of the distribution. HP = high-predictability; LP = low-predictability; $N = 165$.

two nested models predicting recognition accuracy—one that included only type (HP or LP) as a fixed effect, and one that included both type and condition (*audio-only*, *static*, *signal*, *yoked*) as fixed effects. For all models, participants and items were entered as random effects, and the maximal random effects structure justified by the design was used (Barr, Levy, Scheepers, & Tily, 2013; see Supplementary Materials for a description of the random effects structure we employed for each set of analyses). A likelihood ratio test provided by the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017) indicated that a model with type as the only fixed effect was preferred to a model with both type and condition as fixed effects for the analysis of all words ($\chi^2_3 = 3.06$; $p = .38$) as well as the analysis of final words only ($\chi^2_3 = 1.49$; $p = .68$); that is, we found that the circle did not affect recognition in either analysis (see Fig. 1). We performed two additional model comparisons for the sentence-final word data to assess the influence of type (HP vs. LP), as well as the interaction between condition and type. We did not conduct these analyses for the full sentence data, as only the final word was predictable from context. A likelihood ratio test indicated that a model with both condition and type was preferred to a model with only condition ($\chi^2_1 = 31.54$; $p < .001$), suggesting that the effect of type was significant. Examination of the summary output for the full model indicated that HP words were recognized more accurately than LP words ($\beta = -1.11$, $SE = .19$, $z = -5.93$, $p < .001$). Finally, we found that a model without the Condition × Type interaction was preferred to a model that included the interaction ($\chi^2_3 = 3.57$; $p = .31$), indicating that the effect of condition was similar for HP and LP words.

The finding that the abstract visual stimulus used in this study did not facilitate speech recognition is consistent with the results of Schwartz et al. (2004) and Summerfield (1979), and may suggest that some level of phonetic detail is necessary for visual enhancement. However, it is possible that temporal features of the abstract visual stimulus enhanced low-level attentional processes, thereby reducing "listening effort" (LE)—the cognitive resources necessary to comprehend speech (Downs, 1982; see also Pichora-Fuller et al., 2016). If participants were already attending to the speech task to the best of their abilities, then these attentional benefits would not lead to improved recognition, but may instead make the recognition task less cognitively demanding.

Research on LE is based on the assumption that an individual's pool of cognitive and attentional resources is finite (Kahneman, 1973; Rabbitt, 1968), so as a listening task becomes more difficult, fewer resources remain available to complete other tasks simultaneously. Critically, LE levels cannot necessarily be inferred from recognition scores—some interventions, such as noise-reduction algorithms in hearing aids, may reduce LE without affecting speech recognition (Sarampalis, Kalluri, Edwards, & Hafter, 2009). Thus, it may be that an abstract visual stimulus like a modulating circle reduces LE without improving recognition accuracy. Experiment 2 examined this possibility using a dual-task paradigm, a commonly used method of quantifying LE (see Gagné, Besser, & Lemke, 2017).

Experiment 2

Method

Participants

Ninety-six participants, ages 18–28 years, from the Carleton College community participated in Experiment 2. This sample size was predetermined using power analysis, and this experiment was preregistered via the Open Science Framework (https://osf.io/b94yx/). Although we report data from 96 participants, we collected data from 104 individuals and excluded a total of eight from the primary analyses (see Supplementary Materials for more details regarding the power analysis, and the link above to view the preregistered exclusion criteria and to access all stimuli, raw data, and code). Carleton College's Institutional Review Board approved all research procedures. Participants were compensated $5 for 30 minutes of participation.

Stimuli

Experiment 2 employed the semantic dual-task (SDT; Picou & Ricketts, 2014; Strand, Brown, Merchant, Brown, & Smith, in press), in which participants listen to a stream of words and determine as quickly and accurately as possible whether each word is a noun. Speech stimuli consisted of 400 words that were selected from a subset of the SUBTLEX-US database (Brysbaert, New, & Keuleers, 2012), excluding articles and conjunctions, uncommon words (log-frequencies less than three), and long words (more than two syllables or five phonemes). To be consistent with prior research using the SDT (Picou & Ricketts, 2014), 55% of words were predominantly classified as nouns (according to the SUBTLEX-US part of speech dominance data; Brysbaert et al., 2012). The 400 words were divided into four lists that maintained the 55% noun composition, and each list was used in each of the four conditions an equal number of times. Visual stimuli were presented on a 21.5-inch iMac computer via SuperLab 5 (Cedrus), and auditory stimuli were produced by the same female speaker as in Experiment 1, presented in two-talker babble at an SNR of −4 dB. We used QuickTime screen recording to create videos from the output of the custom JavaScript program so that we could collect reaction time data with SuperLab.

Design and procedure

We opted to include only the *audio-only* and signal conditions from Experiment 1 to shorten the experiment and enable a within-subjects design. Participants first completed two recognition-only blocks (*audio-only* and *signal*, order counterbalanced across participants) in which they were asked to repeat the words aloud as they were presented. These blocks were completed without the noun-judgment task and were included to replicate Experiment 1 with words rather than sentences. Next, participants completed two SDT blocks (*audio-only* + *SDT* and *signal* + *SDT*, order counterbalanced across participants).

During the SDT blocks, participants were asked to listen to a stream of words and press a button on a button box (Cedrus RB-740) as quickly and accurately as possible whenever the word was a noun. After making the noun judgment, participants were asked to repeat aloud the word they perceived, regardless of its part of speech. Reaction times to trials in which participants reported perceiving a noun were taken as a measure of LE. Accuracy for the noun classification task was not scored because approximately 84% of nouns can be classified as other parts of speech (Picou & Ricketts, 2014), and because individuals may differ in their ability to classify nouns (see Picou & Ricketts, 2014). In all blocks, the interstimulus interval varied randomly between 2,000 ms and 3,000 ms in 500 ms steps. Participants completed four practice trials before each of the single-task conditions, and eight practice trials before each of the dual-task conditions. Accuracy for the speech recognition task was scored off-line by research assistants. As in Experiment 1, we also collected subjective ratings of task difficulty (see Supplementary Materials).

Results and discussion

Word recognition analysis

Unless otherwise specified, the analyses here followed the conventions of Experiment 1, and details of the random effects structure we employed are available in the Supplementary Materials. The word recognition analysis was performed exclusively on single-task trials. To determine whether condition (*audio-only* vs. *signal*) affected recognition accuracy, we built two nested models predicting accuracy—a full model with condition as a fixed effect and participants and items as random effects, and a reduced model that lacked any fixed effects but was identical to the full model in all other respects. A likelihood ratio test indicated that the reduced model was preferred (χ^2_1 = .01; p = .93). These results replicate those of Experiment 1 using words rather than sentences, and suggest that a modulating circle does not facilitate word recognition; indeed the mean accuracy in the *audio-only* condition (81%, *SD* = 8%) was nearly identical to that in the signal condition (82%, *SD* = 7%).

Listening effort analysis

The LE analysis was performed on the *audio-only* + *SDT* and *signal* + *SDT* trials. As above, we built two nested models, but in this analysis the dependent variable was reaction time to the noun-judgment task. In the full model, condition was entered as a fixed effect, and participants and items were entered as random effects. The reduced model had only random effects. A likelihood ratio test indicated that the larger model was preferred (χ^2_1 = 138.96; p < .001), suggesting that reaction times differed as a function of condition. Examination of the summary output for the full model indicated that reaction times were on average an estimated 185 ms faster in the signal

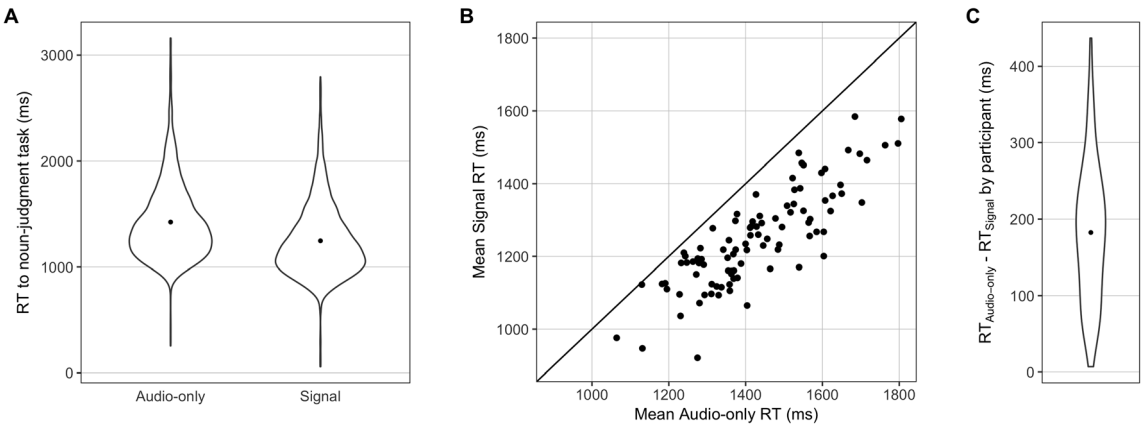

Fig. 2 **a** Violin plots showing RT by condition. Each plot contains all trials during which participants reported perceiving a noun. **b** Scatterplot showing average RTs for each participant in the signal and *audio-only* conditions; the fact that all points are below the line $y = x$ indicates that all participants had faster average RTs in the signal than *audio-only* condition. **c** The difference between average RT in the *audio-only* and signal conditions for each participant. RT = reaction time; ms = milliseconds. N = 96.

condition (β = −185.11, *SE* = 13.51, *t* = −13.70, *p* < .001; *d* = .44; see Fig. 2a; see Supplementary Materials for details on how effect size was calculated). Taken together, the results of Experiment 2 suggest that although the modulating circle does not improve spoken word recognition, it facilitates speeded judgments about the speech. Strikingly, the modulating circle reduced LE on the individual level—every one of the 96 participants in Experiment 2 had faster average reaction times in the signal condition compared to the *audio-only* condition (see Fig. 2b), and nearly 80% (75 of the 96 participants) had reaction time differences larger than 100 ms (see Fig. 2c). Thus, the observed effect appears to be large enough to benefit individual listeners.

General discussion

In these experiments, we demonstrated that an abstract visual signal lacking fine phonetic detail does not improve intelligibility of sentences or words, but it significantly reduces the amount of effort required to comprehend the speech. Although listeners could complete the speech recognition task in Experiment 2 with relatively high accuracy, the abstract visual stimulus made the task easier, which liberated cognitive resources that could be allocated to the secondary task. These results provide further evidence for a dissociation between speech recognition accuracy and LE, corroborating the research showing that hearing-aid noise-reduction algorithms do not improve speech recognition, but do reduce LE (as measured by both faster reaction times and improved recall; Desjardins & Doherty, 2014; Sarampalis et al., 2009). This finding has important clinical implications, as it suggests that hearing assessments that measure recognition accuracy but do not take into account the cognitive requirements of speech understanding may be missing important information about a patient's listening experience. Further, given that these results are analogous to those in the hearing-aid literature, it follows that this research may inform the development of technology aimed at reducing LE and subjectively rated listening difficulty (Bentler, Wu, Kettel, & Hurtig, 2008) rather than improving recognition. That is, the widespread use of noise-reduction algorithms—despite several empirical demonstrations of their ineffectiveness at improving intelligibility—suggests that their cognitive benefits are discernable to the listener, so the development of a similar device that uses a modulating circle may be a fruitful avenue for clinical research.

Such a device would be particularly beneficial for individuals who must expend considerable effort to recognize speech, including listeners who have difficulty hearing over the telephone, and individuals who listen to the radio or audiobooks in noisy environments. Although additional phonetic detail provided by a talking face may greatly improve speech recognition in these situations, this information is not always available. In these circumstances, having access to a modulating circle would allow the listener to expend fewer resources to recognize the speech while attaining the same level of performance, which would improve the listener's ability to successfully retain the information they heard. Indeed, research has demonstrated improved comprehension (Arnold & Hill, 2001), recall (Rabbitt, 1968; Sommers & Phelps, 2016), and recognition memory (Van Engen, Chandrasekaran, & Smiljanic, 2012) when the listening task is less cognitively demanding.

This study also has important implications for research addressing whether adding a visual signal reduces LE or incurs additional processing costs. Conflicting findings in the literature (Gosselin & Gagné, 2011; Mishra, Lunner, Stenfelt, Rönnberg, & Rudner, 2013; Sommers & Phelps, 2016) might be attributable to the paradigm used to quantify LE (e.g., recall vs. dual task), task difficulty, or methodological decisions about whether SNR or performance is equated across *audio-only* and audiovisual conditions. The present paradigm is unique in that both SNR and performance were equated, allowing for a clear interpretation of results, unencumbered by differences in recognition scores.

Another reason for previous conflicting findings about the influence of visual information on LE may be that different features of the visual signal have different effects on LE. That is, low-level features of the visual signal, like the attention-capturing features utilized in the present study, may reduce LE, but extracting fine phonetic detail from a talking face may increase LE. The relative salience of each of these factors may determine whether the overall effect of adding the visual modality in each of these studies is a reduction or an increase in LE. Future research should aim to identify which particular features of the visual signal reduce LE and which, if any, induce additional processing costs. Additional work is also needed to determine whether the benefits shown here with isolated words extend to more naturalistic stimuli such as sentences, or to more difficult listening situations. The current study fills a gap in the audiovisual integration and LE literatures by demonstrating that low-level features of the visual signal reduce LE, possibly by attracting the listener's attention to salient moments in the acoustic input.

Author note Carleton College supported this work. We are grateful to Hunter Brown, Naseem Dillman-Hasso, Lydia Ding, Kate Finstuen-Magro, Alexander Frieden, Maryam Hedayati, Sasha Mayn, Madeleine Merchant, Lucia Ray, Julia Smith, Hettie Stern, Janna Wennberg, and Annie Zanger for assistance with data collection, Xinyu Song for creation of the custom stimulus delivery software, Daniel Hernández for input on experiment design, Adam Putnam for comments on an earlier draft, and Aaron Swoboda for suggestions about data visualization.

References

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology, 92*, 339–355.

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). doi: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 [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://github.com/lme4/lme4/

Bentler, R., Wu, Y.-H., Kettel, J., & Hurtig, R. (2008). Digital noise reduction: Outcomes from laboratory and field studies. *International Journal of Audiology, 47*(8), 447–460.

Bernstein, L. E., Auer, E. T., Jr., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, *44*(1/4), 5–18.

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.

Desjardins, J. L., & Doherty, K. A. (2014). The effect of hearing aid noise reduction on listening effort in hearing-impaired adults. *Ear and Hearing, 35*(6), 600–610.

Downs, D. W. (1982). Effects of hearing aid use on speech discrimination and listening effort. *The Journal of Speech and Hearing Disorders, 47*(2), 189–193.

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.

Gagné, J.-P., Besser, J., & Lemke, U. (2017). Behavioral assessment of listening effort using a dual-task paradigm: A review. *Trends in Hearing*, *21*. doi:https://doi.org/10.1177/2331216516687287

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., van Wassenhove, V., & Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory–visual (cross-modal) synchrony. *Speech Communication*, *44*(1/4), 43–53.

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.

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*, 842–849.

Jordan, T. R., & Sergeant, P. (2000). Effects of distance on visual and audiovisual speech recognition. *Language and Speech, 43*(1), 107–124.

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

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America, 61*(5), 1337–1351.

Kim, J., & Davis, C. (2004). Investigating the audio–visual speech detection advantage. *Speech Communication*, *44*(1/4), 19–30.

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26.

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

Munhall, K. G., Kroos, C., Jozan, G., & Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perception & Psychophysics, 66*(4), 574–583.

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.

Schwartz, J.-L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition, 93*(2), B69–B78.

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., Merchant, M. M., Brown, H. E., & Smith, J. (in press). Measuring listening effort: Convergent validity, sensitivity, and links with cognitive and personality measures. *Journal of Speech, Language, and Hearing Research: JSLHR*.

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.

Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica, 36*, 314–331.

Tye-Murray, N., Spehar, B., Myerson, J., Sommers, M. S., & Hale, S. (2011). Cross-modal enhancement of speech detection in young and older adults: Does signal content matter? *Ear and Hearing, 32*(5), 650–655.

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.