Talking Points: A Modulating Circle Increases Listening Effort Without Improving Speech Recognition in Young Adults

Julia F. Strand, Violet A. Brown & Dennis L. Barbour

Psychonomic Bulletin & Review

ISSN 1069-9384
Volume 27
Number 3

Psychon Bull Rev (2020) 27:536-543
DOI 10.3758/s13423-020-01713-y

Springer

Psychonomic Bulletin & Review (2020) 27:536–543
https://doi.org/10.3758/s13423-020-01713-y

BRIEF REPORT

Talking Points: A Modulating Circle Increases Listening Effort Without Improving Speech Recognition in Young Adults

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

Published online: 3 March 2020

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, Kroos, Jozan, & Vatikiotis-Bateson in *Perception & Psychophysics, 66*(4), 574–583, 2004), and static or moving shapes can improve speech detection accuracy (Bernstein, Auer, & Takayanagi 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). Further, contrary to our hypothesis, the modulating circle increased listening effort despite subjective reports that it made the word recognition task seem easier to complete (Experiment 2). These results suggest that audiovisual speech processing, even when the visual stimulus only conveys temporal information about the acoustic signal, may be a cognitively demanding process.

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"

Note: This paper is a corrected version of a previous manuscript published in 2018 (https://link.springer.com/article/10.3758/s13423-018-1489-7). While attempting to replicate and extend the original work, the first author discovered an error in the stimulus presentation program that invalidated the results. This paper presents the corrected results.

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.

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13423-020-01713-y) 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

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 et al., 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 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 et al., 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 non-face 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 aged 18–23 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 en-

hancement 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 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 Supplemental 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 lowpass 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 acoustic signal amplitude 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 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 (1500 ms–3000 ms in 500 ms steps) after the end of the previous trial.

Responses were scored offline 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 high-predictability words more than low-predictability 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 "On a scale from 1 to 7, how difficult did you find this task?" and "What percentage of the sentences do you think you identified accurately?" These measures were included to assess whether participants' subjective experience of difficulty was affected by the circle.

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

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

was excluded from all analyses due to low accuracy (worse than three SDs 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), and we used the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017) to obtain *p*-values for model parameter estimates. To determine whether condition affected accuracy, we first built 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). Given that the data were binomially distributed (1 = correct; 0 = incorrect), we used generalized linear mixed effects models with a logit link function for this set of analyses. A likelihood ratio test 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 (X^2_3 = 3.06; p = 0.38) as well as the analysis of final words only (X^2_3 = 1.49; p = 0.68); that is, we found that the circle did not affect recognition in either analysis (Figure 1). We performed two additional model comparisons for the sentence-final word data to assess the influence of type (HP versus 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 (X^2_1 = 31.54; p < 0.001), 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 (β = -1.11, SE = 0.19, z = -5.93, p < 0.001). Finally, we found that a model without the condition-by-type interaction was preferred to a model that included the interaction, (X^2_3 = 3.57; p = 0.31), indicating that the effect of condition was similar for HP and LP words.

Five participants failed to complete the subjective effort portion of the task, so N = 160 for the effort analysis. Subjective data were analyzed by comparing ordinary linear regression models, since each participant only responded once. Models predicting participants' subjective ratings of difficulty did not provide a better fit for the data than ones that did not include it, either for judgments of numbers of words correctly identified ($F_{3,159}$ = 0.61, p = 0.61), or for difficulty ($F_{3,159}$ = 0.08, p = 0.97), indicating that subjective measures of difficulty did not differ across participant groups (see Table S1 for group means).

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

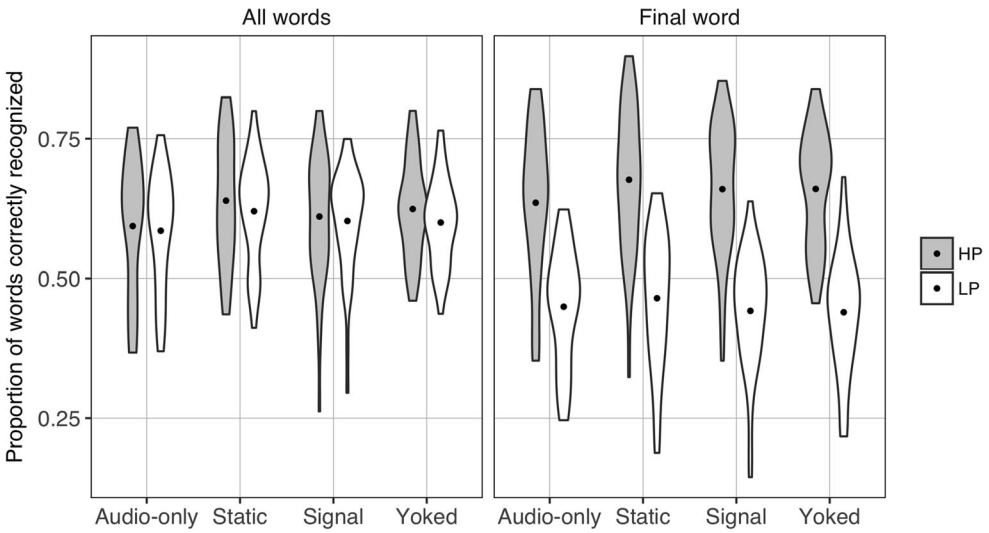

Figure 1. Violin plots showing the distribution of participant mean accuracies by condition and type for the analysis of all words (left) and sentence-final words (right). The dot shows the mean value in each condition, and the width depicts the density of the distribution. HP = high predictability;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

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

Data from 96 participants aged 18–28 from the Carleton College community are included in Experiment 2. This sample size was predetermined using power analysis, and this experiment was pre-registered 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 Supplemental Materials for more details regarding the power analysis, and the link above to view the pre-registered 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, 2018), 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 2000 ms and 3000 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 offline by research assistants. At the end of the study, participants were asked to subjectively rate whether the movement of the circle made it easier to understand the speech using the following prompt: "In this experiment, the dot on the screen sometimes moved and sometimes was still. Did the movement of the dot affect how difficult it seemed to understand the speech?" Participants were given the option of responding "Yes, the movement of the dot made the task seem easier," "Yes, the movement of the dot made the task seem harder," and "No, the movement of the dot did not seem to affect the difficulty of the task."

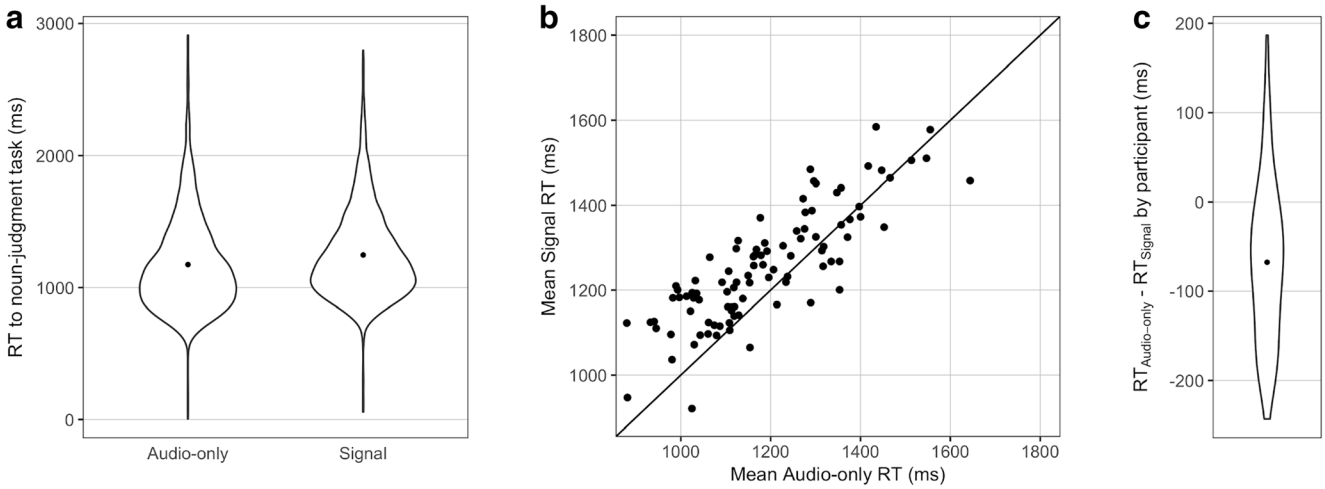

Figure 2. A: Violin plots showing RT by condition. Each plot contains all trials during which participants reported perceiving a noun. B: Scatterplot showing average RTs for each participant in the signal and audio-only conditions; the fact that all points are below the line y = x indicates that all 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.

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 Supplemental Materials. The word recognition analysis was performed exclusively on single-task trials. To determine whether condition (*audio-only* versus *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 ($X^2_1 = 0.01$; $p = 0.93$). 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 ($X^2_1 = 22.03$; $p < 0.001$), 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 65 ms slower in the *signal* condition ($\beta = 64.89$, *SE* = 13.51, $t = 4.80$, $p < 0.001$; Figure 2A).

Subjective Effort Analysis

Two participants failed to respond to the effort question, so the effort question was based on data from 94 participants. A chi-squared goodness-of-fit test indicated that the observed counts of each of the three effort ratings significantly differed from what would be expected by chance (i.e., uniform probability of $^1/_3$; $X^2_2 = 58.11$, $p < .0001$). 66 participants indicated that the circle made it easier, 11 indicated that the circle made it harder, and 17 reported that it made no difference. These results differ from those of Experiment 1, which showed no differences in self-reported difficulty across conditions. The difference between the results of Experiment 1 and Experiment 2 may be a function of study design. That is, it is possible that in a between-subjects study like Experiment 1, participants are asked to rate effort without a clear comparison point, so they resort to reporting their perceived accuracy. Indeed, subjective measures tend to correlate with accuracy rather than objectively measured effort (Seeman & Sims, 2015). In contrast, in within-subjects studies like Experiment 2, participants can more accurately assess task difficulty by comparing perceived effort across conditions.

General discussion

These findings indicate that, at least for the population of normal-hearing young adults used and difficulty level employed here, an abstract visual stimulus did not improve word recognition accuracy. Future work should

evaluate whether the temporal cues provided by the modulating circle may be able to facilitate word recognition under more challenging listening conditions. Word recognition accuracy in Experiment 2 was relatively high (above 80%) and although there was some temporal jitter in when stimuli were presented, the timing was still relatively predictable. Therefore, it is possible that with greater uncertainty about the timing of speech or more difficult listening situations, the temporal cues from the modulating circle may aid recognition. Indeed, recent work using the same stimuli as Experiment 2 but testing older adults with typical age-related hearing loss demonstrated a small but significant improvement in word recognition with the addition of the visual stimulus in that population (Brown, Strand, & Van Engen, in prep).

The results also reveal that although participants tended to subjectively report that the modulating circle made the task seem easier, it actually slowed reaction times to the noun-judgment task. This finding is in line with other work demonstrating inconsistencies across measures of listening effort; although subjective reports and dual-task measures are both assumed to be measuring the same underlying construct, they often generate different patterns of results (see Alhanbali, Dawes, Millman, & Munro, 2019; Strand et al., 2018). Thus, this area of the literature would benefit from additional methodological work to help explicate what each of these measures is capturing.

We had hypothesized that the modulating circle would decrease listening effort by reducing temporal uncertainty and alerting the listeners' attention to salient moments in the speech stream, but instead found that the modulating circle slowed reaction times to the secondary task, indicating increased listening effort. One explanation for this finding is that the modulating circle simply created a distraction for participants. The detrimental effects of trying to complete multiple tasks simultaneously are well established (see Koch, Poljac, Müller, & Kiesel, 2018), and it may be that the concurrent presentation of auditory and visual stimuli resulted in an attentional bottleneck that slowed reaction times. Alternatively, the additional load may come from attentional costs associated with distraction or the process of audiovisual integration. The results reported here are consistent with prior work showing that the presence of a talking face also increases listening effort (see Brown & Strand, 2019; Gosselin & Gagné, 2011). In the case of a talking face, that additional cost may be offset by the beneficial phonetic information about speech that a face provides, such as cues to place of articulation. Here, however, the circle does not provide enough detail to render recognition benefits but still incurs the two-channel processing cost

References

Alhanbali, S., Dawes, P., Millman, R. E., & Munro, K. J. (2019). Measures of listening effort are multidimensional. *Ear and Hearing*. https://doi.org/10.1097/AUD.0000000000000697

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/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."* R foundation for statistical computing, Vienna, 12. Retrieved from https://github.com/lme4/lme4/

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.

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

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.

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*, 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., Wassenhove, V. van, & 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.pdf. *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.

Koch, I., Poljac, E., Müller, H., & Kiesel, A. (2018). Cognitive structure, flexibility, and plasticity in human multitasking—An integrative review of dual-task and task-switching research. *Psychological Bulletin*, 144(6), 557–583. https://doi.org/10.1037/bul0000144

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, Articles*, *82*(13), 1–26.

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.

Seeman, S., & Sims, R. (2015). Comparison of psychophysiological and dual-task measures of listening effort. Journal of Speech, Language, and Hearing Research, 58(6), 1781–1792. https://doi.org/10.1044/2015_JSLHR-H-14-0180

Strand, J. F., Brown, V. A., Merchant, M. M., 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*.

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., 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.