Attention, Perception, & Psychophysics (2019) 81:1743–1756
https://doi.org/10.3758/s13414-019-01772-x

REGISTERED REPORTS AND REPLICATIONS

"Paying" attention to audiovisual speech: Do incongruent stimuli incur greater costs?

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

Published online: 13 June 2019

Abstract
The McGurk effect is a multisensory phenomenon in which discrepant auditory and visual speech signals typically result in an illusory percept. McGurk stimuli are often used in studies assessing the attentional requirements of audiovisual integration, but no study has directly compared the costs associated with integrating congruent versus incongruent audiovisual speech. Some evidence suggests that the McGurk effect may not be representative of naturalistic audiovisual speech processing – susceptibility to the McGurk effect is not associated with the ability to derive benefit from the addition of the visual signal, and distinct cortical regions are recruited when processing congruent versus incongruent speech. In two experiments, one using response times to identify congruent and incongruent syllables and one using a dual-task paradigm, we assessed whether congruent and incongruent audiovisual speech incur different attentional costs. We demonstrated that response times to both the speech task (Experiment 1) and a secondary vibrotactile task (Experiment 2) were indistinguishable for congruent compared to incongruent syllables, but McGurk fusions were responded to more quickly than McGurk non-fusions. These results suggest that despite documented differences in how congruent and incongruent stimuli are processed, they do not appear to differ in terms of processing time or effort, at least in the open-set task speech task used here. However, responses that result in McGurk fusions are processed more quickly than those that result in non-fusions, though attentional cost is comparable for the two response types.

Keywords McGurk effect · Audiovisual integration · Dual-task · Listening effort · Response time

Introduction

Multisensory integration is a hallmark of perceptual processing – evidence for cross-modal interactions has accumulated for nearly every combination of senses (Gottfried & Dolan, 2003; Lackner, 1977; Shankar, Levitan, Prescott, & Spence, 2009; Zampini & Spence, 2004), but one of the most widely studied multisensory interactions is that between audition and vision (e.g., McGurk & MacDonald, 1976). Audiovisual (AV) integration has been demonstrated using both speech (Erber, 1969; McGurk & MacDonald, 1976; Sumby & Pollack, 1954) and non-speech (Saldaña & Rosenblum, 1993) stimuli. Within the speech literature, the interactions between audition and vision have been studied primarily in two ways: (1) Using congruent stimuli, in which the auditory and visual modalities present the same speech input, and (2) using incongruent stimuli, in which the two modalities present mismatched speech inputs (e.g., hearing "ba" and seeing "ga"; Brancazio, 2004; McGurk & MacDonald, 1976). The former paradigm has demonstrated that recognition accuracy is improved when listeners can see and hear the talker relative to hearing alone (Erber, 1969; Grant, Walden, & Seitz, 1998; Sumby & Pollack, 1954; Van Engen, Phelps, Smiljanic, & Chandrasekaran, 2014). The latter often results in the perception of a syllable or word that represents a fusion of features from the two modalities (the "McGurk effect"; McGurk & MacDonald, 1976).

Despite a large body of research, fundamental questions persist about the process of integrating speech information from two modalities. For instance, the literature is mixed on whether integration is a distinct stage of AV speech processing or is instead simply a consequence of the speech recognition process (Tye-Murray, Spehar, Myerson, Hale, & Sommers,

✉ Violet A. Brown
violet.brown@wustl.edu

✉ Julia F. Strand
jstrand@carleton.edu

[1] Psychological and Brain Sciences, Washington University in St. Louis, St. Louis, MO, USA

[2] Carleton College, Northfield, MN, USA

2016). In this paper, we use the term "integration" to refer to combining auditory and visual cues into a unified percept, but do not advocate that this must be achieved via a distinct stage or mechanism. Indeed, instances in which a participant appears to not make use of both the auditory and visual signals may reflect a failure in a dedicated stage of multimodal integration, but may also be attributed to problems with assigning percepts to phoneme categories, being distracted, recruiting additional brain areas to detect and resolve conflict, or something else. This paper does not attempt to dissociate among these possibilities, but instead aims to clarify a second outstanding issue in the integration literature: whether integrating speech information from two modalities occurs automatically or requires attentional resources.

Some work indicates that AV integration occurs automatically (Colin et al., 2002; Soto-Faraco, Navarra, & Alsius, 2004), whereas some suggests that integration is attentionally demanding (Alsius, Möttönen, Sams, Soto-Faraco, & Tiippana, 2014; Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Alsius, Navarra, & Soto-Faraco, 2007; Talsma & Woldorff, 2005; Tiippana, Andersen, & Sams, 2004; see also Van der Burg, Brederoo, Nieuwenstein, Theeuwes, & Olivers, 2010, for evidence that the cross-modal integration of semantic information requires attention). These conflicting results may be due in part to differences in stimulus materials across studies – some experiments within the AV integration literature use congruent speech stimuli (Mishra, Lunner, Stenfelt, Rönnberg, & Rudner, 2013a, 2013b; Sommers & Phelps, 2016) and others use incongruent (McGurk) speech stimuli (Alsius et al., 2014, 2005, 2007; Colin et al., 2002; Soto-Faraco et al., 2004; Tiippana et al., 2004; Tuomainen, Andersen, Tiippana, & Sams, 2005). Although these studies all aim to assess the attentional requirements of AV integration, no study has directly compared the costs associated with integrating congruent versus incongruent speech.[1] Thus, it remains unclear whether integrating incongruent stimuli incurs different processing costs than integrating congruent stimuli, and consequently, whether these two types of stimuli can be used interchangeably in studies assessing the attentional requirements of AV integration.

Individual differences in susceptibility to the McGurk effect are commonly used as a measure of general AV integration ability, but there is evidence that McGurk susceptibility may not accurately reflect the ability to integrate more naturalistic AV speech. Grant and Seitz (1998) found that despite the correlation between McGurk susceptibility and the ability to benefit from the addition of visual speech information ("visual benefit"), McGurk susceptibility did not contribute significant unique variance to regression models predicting visual benefit once lip-reading ability and other measures of integration ability were controlled for. Further, Van Engen et al. (2017) recently found that individual differences in McGurk susceptibility and visual benefit were unrelated to one another. Although these studies do not directly address the attentional requirements of AV integration, they suggest that integrating congruent and incongruent AV speech may rely on different underlying mechanisms.

A growing body of fMRI evidence has also demonstrated a dissociation between the processes involved in integrating congruent and incongruent AV speech. Calvert and colleagues (Calvert, Campbell, & Brammer, 2000) showed that the left superior temporal sulcus (STS) exhibits supra-additive activity in response to congruent AV speech and sub-additive activity in response to incongruent AV speech (relative to the sum of the unimodal responses; note, however, that this study used mismatched auditory and visual sentences rather than McGurk stimuli). Further, Erickson et al. (2014) showed that the left posterior STS exhibits greater activity in response to congruent AV speech compared to auditory-only and visual-only speech, but that a different region, the left posterior superior temporal gyrus (pSTG), exhibits greater activity in response to incongruent speech, suggesting a dissociation in the cortical regions recruited during congruent versus incongruent AV speech processing. Additionally, Morís Fernández et al. (2017) recently demonstrated that compared to congruent AV speech, incongruent speech results in more activity in brain regions associated with general conflict processing (the anterior cingulate cortex), as well as speech-specific conflict processing (the inferior frontal gyrus). Taken together, these results illustrate neuroanatomical differences in how congruent and incongruent AV speech are processed, and suggest that integrating incongruent speech requires the recruitment of additional brain areas to detect and resolve the conflict.

To date, the only behavioral evidence that suggests that integrating incongruent speech is more resource intensive than integrating congruent speech is the finding that response times to incongruent stimuli tend to be slower than those to congruent stimuli (Beauchamp, Nath, & Pasalar, 2010; Green & Kuhl, 1991; Keane, Rosenthal, Chun, & Shams, 2010; Massaro & Cohen, 1983; Nahorna, Berthommier, & Schwartz, 2015; Norrix, Plante, & Vance, 2006; Tiippana, Puharinen, Möttönen, & Sams, 2011). However, two notable experimental choices complicate interpretation of this finding: collapsing across response types to incongruent stimuli (an analytical decision) and using a closed-set task (a methodological decision). The former decision may lead to spurious results because trials in which participants demonstrate the McGurk effect (referred to here as "fusion trials") and those in which they do not (referred to here as "non-fusion trials") may reflect differences in perceptual and cognitive processing, and the latter is problematic because the task constraints placed on participants during closed-set paradigms may be

[1] Although the term "incongruent" is sometimes used to refer to non-illusory mismatched AV stimuli (e.g., $A_gV_b = AV_{gb}$), here the term refers to mismatched AV stimuli that commonly result in a McGurk fusion.

expected to differentially affect response times to congruent and incongruent stimuli.

There are multiple reasons that fusion and non-fusion trials might be processed at different speeds. Non-fusion trials could be faster than fusion trials for at least three reasons. First, integration may have failed to occur (or occurred to a lesser extent) during non-fusion trials, so a perceptual decision was made quickly without an additional integration stage (i.e., there was one less processing step in non-fusion compared to fusion trials; note that this interpretation assumes that integration requires effort, or at least additional processing time, a claim for which evidence is equivocal; Gosselin & Gagné, 2011a; Sommers & Phelps, 2016). Second, the participant may have closed their eyes or completely ignored the visual signal during non-fusion trials, and therefore the perceptual system effectively only processed the auditory stream (again, this relies on the assumption that some component of the integration process requires processing time or effort). Third, the percept that arose during fusion trials may not have cleanly fit into a perceptual category (Brancazio, 2004; Massaro & Ferguson, 1993), so assigning this imperfect representation to a category required additional processing time or resources that exceeded those required for processing non-fusion trials.

Alternatively, fusion trials may be processed more quickly than non-fusion trials for multiple reasons. In non-fusion trials, participants may have tried to integrate the auditory and visual information, and upon failure to do so, had to resort to processing the auditory input and assigning it to a perceptual category, which required additional time or resources beyond those required for processing trials during which integration successfully occurred (i.e., there is an extra step in processing non-fusion trials). Another possible explanation for faster response times to fusion trials is that participants may have noticed the incongruity more often in non-fusion trials, and the recruitment of conflict detection and resolution mechanisms – or simply distraction – slowed responding (but note that awareness of the discrepancy between the auditory and visual signals does not necessarily mean that participants failed to experience the McGurk effect; Soto-Faraco & Alsius, 2007; Strand, Cooperman, Rowe, & Simenstad, 2014). Finally, the influence of the visual signal on the non-fusion trials may have been strong enough that the resulting percept, though ultimately classified as a non-fusion, was such a poor category exemplar that assigning it to an phoneme category slowed responses (Brancazio & Miller, 2005; Gentilucci & Cattaneo, 2005). Given the numerous reasons that response times to fusion and non-fusion trials may differ, combining the response types when comparing incongruent stimuli to congruent stimuli clouds the ability to make inferences about the processing costs of integration.

An additional complication to interpreting the finding that response times to incongruent stimuli are slower than those to congruent stimuli is the fact that all of the studies reporting this finding recorded response times using a button or key press in a forced-choice task. Unlike open-set tasks, in which participants repeat aloud what they perceive, closed-set tasks limit response alternatives to one of several possibilities. Closed-set tasks can lead to different patterns of results than open-set tasks (Alsius, Paré, & Munhall, 2017; Clopper, Pisoni, & Tierney, 2006; Colin, Radeau, & Deltenre, 2005), and in this case may inflate differences in response times to congruent and incongruent speech. For instance, consider a case where participants are presented with one type of incongruent stimulus (e.g., auditory /b/ paired with visual /g/, denoted as A_bV_g) and one of three types of congruent stimuli, corresponding to the auditory and visual components of the incongruent stimulus and the expected fusion (A_bV_b, A_gV_g, A_dV_d). In one condition, participants must press one of three buttons (corresponding to /bɑ/, /gɑ/, or /dɑ/) as quickly as possible indicating what they perceived, and in the other, they must respond as quickly as possible by repeating aloud what they perceived. Given that McGurk percepts tend to be poor exemplars for phoneme categories (Brancazio, 2004; Massaro & Ferguson, 1993; Rosenblum & Saldaña, 1992), and non-fusion responses can still show visual influence in McGurk experiments (Brancazio & Miller, 2005; Gentilucci & Cattaneo, 2005), the percept associated with A_bV_g is not likely to fit neatly into any phoneme category. Thus, responses to the incongruent stimuli may be particularly slowed in the closed-set condition (which requires the additional task of assigning the percept to a discrete perceptual category and determining category goodness) compared to the open-set condition, in which participants are free to produce whatever they perceived.

Critically, the additional demands of a closed-set task for incongruent stimuli may not transfer to congruent stimuli which are, at least in laboratory settings, near-perfect exemplars of phoneme categories (see Massaro & Cohen, 1983, for evidence that AV trials with ambiguous auditory components are accurately identified according to the visual input at high rates). In fact, responses to congruent stimuli may be speeded in closed-set compared to open-set tasks; if congruent AV stimuli provide a strong match for a phoneme category that has been defined by a response key, this may facilitate recognition of the spoken phoneme and speed responses. For example, deciding that a congruent AV /dɑ/ is a /dɑ/ rather than a /bɑ/ may take less time than assigning a congruent AV /dɑ/ to one of many phoneme categories in an open-set task. Thus, the differences in response times to congruent and incongruent AV speech are likely to be larger in the closed-set tasks that are prevalent in the literature than they would be in open-set tasks.

The goal of the present study was to determine whether integrating congruent and incongruent AV speech incurs different attentional costs. Despite the apparent behavioral evidence in favor of this claim, the studies showing that response times are slower to incongruent compared to congruent stimuli are confounded by (1) incongruent trials during which

participants fail to perceive a McGurk fusion, and (2) task constraints in which participants are required to make closed-set categorizations rather than open-set identifications – these analytical and methodological choices preclude a clear interpretation of previous results.

The primary purpose of Experiment 1 was to use an open-set task to determine whether incongruent stimuli that result in McGurk fusions are processed more slowly than congruent stimuli. In Experiment 1, we also collapsed incongruent trials across response types (fusion and non-fusion), following the procedures of prior research, and compared response times to congruent and combined incongruent stimuli. We expected to replicate the previous finding that response time are slower to incongruent compared to congruent stimuli. If we replicate prior research when we collapse across incongruent response types but not when we analyze only fusion responses, this would suggest that the previously reported response time differences between congruent and incongruent trials are driven by the incongruent trials in which participants fail to experience the McGurk effect. Finally, we compared response times to fusion and non-fusion responses to determine whether these two response types are processed differently. Given that several arguments could be made in favor of faster processing of either response type (see above), we do not have a hypothesis about the direction of the effect, or even whether response times to the two response types will indeed differ. We therefore report this exploratory analysis not to provide evidence for or against a specific hypothesis, but rather to provide insight into a possible confound that has not yet been addressed in the literature.

In Experiment 2, we used a dual-task paradigm to directly test whether integrating incongruent AV speech is more attentionally demanding than integrating congruent speech. Dual-task paradigms rely on the assumption that humans possess a limited pool of cognitive resources (Kahneman, 1973; Pashler, 1994), so as the primary task becomes more difficult, fewer resources are available to quickly and accurately complete the secondary task. Thus, slower response times to the secondary task reflect additional attentional costs. Given the possible dissociation between performance on the primary and secondary tasks in a dual-task paradigm, response times to the speech task itself may provide a less conclusive test of the hypothesis that integrating incongruent stimuli requires more cognitive resources than integrating congruent stimuli. For example, several studies have demonstrated that listeners can achieve equivalent performance on a speech recognition task under easy and difficult listening conditions, but still expend more "listening effort" in the more difficult condition (Desjardins & Doherty, 2014; Sarampalis, Kalluri, Edwards, & Hafter, 2009; Strand, Brown, & Barbour, 2018). In this case, response time to the secondary task, and not performance on the speech task itself, reflects attentional demand. To the extent that response time is a measure of performance on the syllable identification task employed in the current study, it may be expected that response times to a secondary task would show a different pattern of results than those to the primary task.

It is also possible that slower response times to incongruent than congruent stimuli simply reflect a more exhaustive search through phoneme representations in order for the perceiver to locate a match to the imperfect speech input, which may require processing time in the absence of cognitive effort. An analogous example of increased processing time without increased resource expenditure comes from semantic verification tasks. One explanation for why participants are faster to indicate that a robin is a bird than a turkey is a bird (Rips & Shoben, 1973) is that when a concept is activated, the activation spreads to related concepts, but the strength of the activation decreases as it spreads outwards (Collins & Loftus, 1975). Therefore, response times to affirm that a robin is a bird are faster because activation is stronger for closely related concepts, not necessarily because affirming that a turkey is a bird is a more effortful process. Similarly, lexical decisions are typically faster to words than non-words because for words, the search terminates as soon as the listener matches the input to an entry in the mental lexicon, but for non-words, the listener must complete an exhaustive search (Caramazza & Brones, 1979; Forster & Bednall, 1976; Gilchrist & Allen, 2015; Rubenstein, Garfield, & Millikan, 1970). This search may proceed automatically, in which case response time to the speech task itself is a poor indicator of attentional cost. Thus, although slower response times to incongruent speech compared to congruent speech may suggest that integrating incongruent speech is more attentionally demanding, measuring response time to an unrelated secondary task is a more direct test of the cognitive requirements associated with integrating congruent versus incongruent speech.

Therefore, Experiment 2 employed a dual-task paradigm, a widely used and well-established paradigm in the speech perception and attention literatures (Alsius et al., 2005, 2007; Downs, 1982; Gagné, Besser, & Lemke, 2017; Gosselin & Gagné, 2011a; Sarampalis et al., 2009; Strand, Brown, Merchant, Brown, & Smith, 2018), to assess the attentional costs associated with integrating congruent and incongruent AV speech. Provided that the secondary task is sufficiently difficult – the attentional requirements of the primary task cannot be assessed if the secondary task can be completed automatically – poorer performance is expected when the primary task exhausts more of the individual's attentional resources (Navarra, Alsius, Soto-Faraco, & Spence, 2010). Prior research suggests that vibrotactile tasks can serve as a sufficiently difficult secondary task to detect differences in the attentional costs of the primary task (Fraser, Gagné, Alepins, & Dubois, 2010; Gosselin & Gagné, 2011a, 2011b). For this reason, participants completed a vibrotactile task while simultaneously completing a speech task with congruent and

incongruent AV speech. An additional benefit of a vibrotactile task (rather than an auditory or visual one) is that any observed effects must be attributable to cognitive effort rather than sensory interference. In Experiment 2, we only analyzed incongruent fusion trials to ensure that the visual signal was actually processed (i.e., we did not collapse across incongruent response types). We hypothesized that response times to the secondary task would be slower in the incongruent fusion condition compared to the congruent condition, indicating that additional cognitive demands were incurred to process mismatched auditory and visual speech relative to congruent speech. This would suggest that findings based on the commonly employed McGurk paradigm may not accurately represent processing of more naturalistic, congruent AV speech (Brancazio & Miller, 2005; Van Engen et al., 2017). We also performed an exploratory analysis akin to that in Experiment 1 to determine whether the attentional costs associated with integrating fusion versus non-fusion responses differ.

To summarize, the goal of Experiment 1 was to determine whether integrating incongruent AV stimuli that result in McGurk fusions requires more processing time than integrating congruent AV speech using an open-set task. We also aimed to replicate previous research by comparing congruent trials to incongruent trials, collapsed across response types. Finally, we report an exploratory analysis aimed at assessing whether the time required to process fusion and non-fusion trials differs. Experiment 2 tests whether the slower response times to incongruent stimuli that have been previously reported (and are hypothesized in Experiment 1) are indeed indicative of greater attentional costs associated with integrating incongruent AV stimuli, or whether this processing instead occurs with minimal attentional demands. We also performed an exploratory analysis aimed at assessing whether fusion and non-fusion responses incur different attentional costs.

In this registered report, the exclusion criteria, analysis plan, and sample sizes were registered and approved by *Attention, Perception, & Psychophysics* on 27 March 2018, prior to data collection. The Stage 1 manuscript that was accepted in principle is available at https://osf.io/8t7an/. Any deviations from the approved report are noted explicitly. All stimuli, data, and code for analyses (which contains details regarding the precise random effects structure we employed and decisions made in cases of non-convergence or singularity) are available at https://osf.io/z6kv3/.

Pilot 1: Ensuring auditory intelligibility

Before conducting the main experiments, we first conducted two pilot experiments to ensure that the auditory components of all congruent and incongruent tokens were highly recognizable (Pilot 1) and to ensure that the incongruent stimuli could effectively elicit the McGurk effect (Pilot 2).

Method

Participants A total of 20 undergraduate participants from Carleton College completed a pilot study to assess the intelligibility of the auditory stimuli to ensure that visually influenced responses on the AV trials were not due to faulty auditory materials. Testing took approximately 30 min, and participants were compensated US$5 for their time. The Carleton College Institutional Review Board approved all research procedures.

Stimuli and procedure All experiments reported in this manuscript were conducted in SuperLab 5 (Cedrus) and administered on a 21.5-in. iMac computer. Auditory stimuli were presented at a comfortable listening level through Seinheisser HD 280 Pro headphones (Pilot 1, Pilot 2, and Experiment 2), or through Beyerdynamic DT 100 headphones with an Aphex HeadPod Model 454 high output headphone amplifier (Experiment 1). When necessary, verbal responses were recorded using Audacity (version 2.1.2) and scored offline by research assistants (Pilot 2, Experiment 1, and Experiment 2).

Video stimuli were recorded using a Panasonic AG-AC90 camera and audio stimuli were recorded with a Shure KSM-32 microphone with a plosive screen by a female speaker. Noise was removed from all auditory tracks before creating speech stimuli, and all auditory stimuli were equalized on root-mean-square (RMS) amplitude using Adobe Audition (version 9.2.0). Pilot 1 consisted of 14 tokens of each of the 12 auditory stimuli that could be used in the main experiment (possible stimuli that could be presented in the main experiment include: /bɑ/, /dɑ/, /fɑ/, /gɑ/, /kɑ/, /mɑ/, /nɑ/, /pɑ/, /tɑ/, /vɑ/, /ðɑ/, and /θɑ/, based on the auditory and visual components of incongruent stimuli, as well as the expected fusions). Each token was presented three times, and the order was randomized, resulting in 504 trials (12 stimuli * 14 tokens * 3 repetitions). After each token, participants entered what they perceived in a textbox. Before beginning the experiment, participants completed three practice trials.

Results

The purpose of the first pilot study was to select the eight tokens with the highest intelligibility for each of the 12 auditory stimuli. Only one of the 12 stimuli ("fa") had fewer than eight tokens with recognition accuracies above 90%. Given that syllables like "fa" tend to be highly confusable in auditory-only settings (Toscano & Allen, 2014), even when the recordings are high quality and the stimuli are presented without background noise, we included two tokens of "fa" that were recognized at rates of 88% and 87%. The top eight tokens for each of the 12 stimuli selected from this pilot study were used in the second pilot study. Thus, all stimuli in the main experiments were recognized at rates of at least 87%, and 66 of the 96 tokens (68.75%) were recognized with 100% accuracy.

Pilot 2: Selecting McGurk stimuli

Given that there exists wide variability in the extent to which incongruent AV stimuli elicit the McGurk effect (Basu Mallick, Magnotti, & Beauchamp, 2015), this pilot study was included to help ensure effective incongruent stimuli.

Method

Participants Twenty-one participants, none of whom had participated in the first pilot study, were recruited from the Carleton College community. One participants' data were not analyzed because they did not complete the task correctly. Testing took approximately 15 min, and participants were compensated US$3 for their time. All procedures were approved by the Carleton College Institutional Review Board.

Stimuli and procedure We first created eight tokens of each of seven incongruent stimuli that have previously been used to elicit the McGurk effect (A_bV_g, A_bV_f, A_mV_g, A_mV_t, A_pV_g, A_pV_k, A_tV_b). Stimuli and expected fusions were determined from Magnotti et al. (2015) and Strand et al. (2014). All stimuli were created in iMovie (version 10.1) by aligning the consonant bursts of the original AV track with the to-be-spliced auditory tracks, then deleting the original auditory track from the video recording (e.g., to create the McGurk stimulus A_bV_g, we took the A_gV_g stimulus, matched the audio track in time with the audio recording of /bɑ/, then deleted the auditory /gɑ/). We created AV stimuli using only the highly recognizable auditory tokens that had been selected from Pilot 1, and ensured that particular auditory and visual tokens were never repeated within a stimulus (e.g., every token of A_bV_g used a different auditory token and a different visual token, but these tokens were repeated across stimuli).

Each AV token was presented four times, and order was randomized, resulting in 224 trials (7 stimuli * 8 tokens * 4 repetitions). Due to experimenter error, only seven of the eight tokens were presented for two McGurk stimuli (A_mV_t and A_pV_g), and one of the eight tokens for two stimuli (A_bV_g and A_mV_g) was repeated instead. That meant that the incongruent stimuli weren't presented at equal rates, but given that every stimulus was presented a minimum of 24 times to each participant, we proceeded with stimulus selection as planned.

After each stimulus, participants were asked to repeat aloud what they perceived (see Table 1 for a list of expected fusions for each McGurk stimulus). Before beginning the experiment, participants completed three practice trials. The four incongruent stimuli that elicited the highest McGurk fusion rates were selected for use in the main experiments, provided that no more than two of the four incongruent stimuli consisted of the same auditory syllable. For each of those four stimuli, we then selected the six tokens that elicited that highest McGurk rates.

Table 1 McGurk stimuli and expected fusions

Auditory stimuli	Visual stimuli	Expected fusions
bɑ	gɑ	dɑ, ðɑ, θɑ
bɑ	fɑ	vɑ
mɑ	gɑ	nɑ
mɑ	tɑ	nɑ
pɑ	gɑ	kɑ
pɑ	kɑ	tɑ, ðɑ, θɑ
tɑ	bɑ	pɑ

Results

The four stimuli with the highest fusion rates were A_bV_f (60.16%), A_pV_k (57.50%), A_bV_g (34.72%), and A_mV_t (16.25%). These rates are comparable to those reported elsewhere (Basu Mallick et al., 2015). Fusion rates for the top six tokens within those four stimuli ranged from 12.50% to 72.50%. These 24 tokens were used in both of the main experiments.

Experiment 1

Method

Participants We collected data from 95 participants (ages 18–30 years) from the Washington University in St. Louis community to ensure we would reach the approved sample size of 85 after applying the approved exclusion criteria, but no participants were excluded on this basis. Technical difficulties precluded analysis of two participants' data (the computer crashed partway through the experiment), and one participant's data could not be analyzed because their voiced responses were unintelligible. We only analyzed data from the first 85 usable data files, meaning data from seven participants were not included. All procedures were approved by the Washington University in St. Louis Institutional Review Board.

Stimuli Stimuli consisted of 24 unique incongruent tokens (six tokens of each of the four stimuli determined by the second pilot study). Congruent stimuli consisted of the auditory and visual components that made up the incongruent stimuli, as well as the expected fusions (see Table 1). For example, for the McGurk stimulus A_bV_g, congruent stimuli included A_bV_b, A_gV_g, A_dV_d, $A_ðV_ð$, and $A_θV_θ$ (for the two stimuli with multiple possible fusions – A_bV_g and A_pV_k – we accepted both /ðɑ/ and /θɑ/ because it is difficult to distinguish between these two responses in audio recordings). This resulted in 11 incongruent stimuli (see *Procedure* section for details

regarding the number of tokens and repetitions for congruent and McGurk stimuli). Whenever possible, congruent stimuli were created with the same auditory and visual speech tokens as the incongruent stimuli to increase the similarity of stimuli across conditions. This was not possible for some congruent stimuli representing the expected fusion of a McGurk stimulus (e.g., one possible fusion for A_bV_g is /ðɑ/, but none of the incongruent stimuli have /ðɑ/ as either the auditory or visual component, so the auditory and visual components of $A_ðV_ð$ were not the same as incongruent trials). Within the congruent condition, each AV syllable was created by combining the auditory and visual components of different recordings of the same syllable, and we only used the highly recognizable auditory tokens that had been identified in the first pilot study. This step was included to ensure that any observed effects could not be attributed to the splicing process itself.

Within each video, the onset of the audio began approximately 275 ms after the onset of the video.[2] After each video ended, the frozen image remained on the screen until participants verbally responded or 2,975 ms had elapsed from the onset of speech (i.e., 3,250 ms after the onset of the video), at which point a white screen with an interstimulus interval (ISI) of 1,000 ms appeared.[3] When participants responded, the video disappeared (to help indicate to them that the voice key had picked up their response).

Procedure Participants were presented with randomly intermixed congruent and incongruent AV syllables and were asked to repeat aloud the syllable they perceived, and to avoid making any unnecessary noise (Luce & Pisoni, 1998). They were instructed that if the voice key did not pick up their auditory response, they should advance to the next trial by pushing the spacebar. They practiced using the voice key by repeating eight written syllables (constituting all the auditory tokens, visual tokens, and likely fusions) prior to beginning the speech task.[4] Response times were recorded from the onset of the video file to the onset of the voiced response using a FiFine Games-K667 condenser microphone connected to a Cedrus SV-1 voice key. The comparison of primary interest to the current study is response time to congruent syllables versus response time to incongruent syllables during which participants perceive a McGurk fusion.

Participants completed eight practice trials with auditory stimuli, during which an experimenter remained in the room to ensure that participants were completing the task correctly. In the main task, participants were presented with 288 total stimuli. Each of the 24 incongruent tokens was presented six times, for a total of 144 incongruent trials (4 stimuli * 6 tokens * 6 repetitions). Because the number of unique congruent stimuli was dictated by the incongruent stimuli determined by the second pilot study, we presented each token enough times to ensure that there were 144 congruent trials. Given that we included 11 unique stimuli, we opted to create four tokens of each stimulus, and each token was repeated three times. This amounted to 132 incongruent trials, so we randomly selected four congruent stimuli ("fa," "ka," "pa," and "tha") and created one additional token of each of them to be repeated three times to ensure that we had 144 total stimuli (11 stimuli * 4 tokens * 3 repetitions + 4 stimuli * 1 token * 3 repetitions).

Results and discussion

At the trial level, we excluded trials with response times longer than 2,000 ms from speech onset (i.e., responses that occurred more than 2,275 ms after speech onset; less than 1% of the data), trials on which the response time was more than three median absolute deviations (MADs) below or above that participant's median response time for that trial type (2.89% of trials), and trials in which the voice key was not triggered, the participant failed to respond, or the response was unclear (2.45% of trials). The reported analyses consisted of 11,664 congruent trials, 5,931 fusion responses, and 5,596 non-fusion responses. By-participant McGurk fusion rates ranged from 9.29% to 94.12%, with a mean fusion rate of 51.45%. The mean fusion rate by stimulus ranged from 33.48% to 82.53%.

Data were analyzed using mixed-effects modeling via the *lme4* package (version 1.1-21; Bates et al., 2014) in R (version 3.5.2; R Core Team, 2016), and, where appropriate, *p*-values from mixed effects models were obtained via the *lmerTest* package (version 3.1-0, Kuznetsova, Brockhoff, & Christensen, 2017). Following the recommendations of Barr, Levy, Scheepers, and Tily (2013), we used the maximal random effects structure justified by the design. In cases of non-convergence, we simplified the random effects structure based on contributions of the variance components to the total variance, and adjusted control parameters in the *lme4* package. For both Experiment 1 and Experiment 2, participants and speech stimuli were entered as random effects. In both experiments, we first compared response times to congruent trials and McGurk fusion responses only. Experiment 1 (but not 2) also compared congruent trials and combined McGurk fusion and non-fusion responses. In these analyses, random slopes

[2] In natural speech, the face movement typically precedes the audio signal. We had originally planned to trim the video files such that the video began 250 ms before audio onset. However, this proved to be too short for some tokens. Starting the video 250 ms before audio onset would have resulted in the visual signal starting with an already open mouth or pursed lips. Thus, we increased the video lead time to approximately 275 to ensure each video file started with a neutral face.

[3] We had originally specified an ISI of 2,000 ms, but changed it to 1,000 ms because it seemed unnecessarily long and to help reduce the total length of the experiment.

[4] These orthographic practice trials were not part of the original manuscript that was accepted in principle. We opted to include them after realizing that it may be helpful for participants to become familiar with the level of vocalization that was necessary to trigger the voice key.

for speech stimuli were not considered because the nature of the McGurk effects warrants a between-items design – that is, each item was either a congruent stimulus or a McGurk stimulus. However, in the final analysis for each experiment we compared response times to McGurk fusion and McGurk non-fusion responses, and we included the by-stimulus random slope for response type (fusion or non-fusion) in the random effects structure because a given McGurk stimulus could be classified as both a fusion and a non-fusion response depending on the participant and trial – that is, for these analyses, response type was within-items.

Congruent versus McGurk fusion analysis The first analysis aimed to determine whether response times to congruent AV trials differed from those to incongruent trials that elicited a McGurk fusion. We first built a full model with stimulus type (congruent vs. McGurk fusion) as a fixed effect and compared this model to a reduced model that lacked stimulus type as a fixed effect but was identical to the full model in all other respects. A likelihood ratio test indicated that the model without the stimulus type provided a better fit for the data ($\chi^2_1 = 0.80$; $p = 0.37$). We therefore did not find evidence supporting our hypothesis that response times to incongruent stimuli that result in a fusion are slower than those to congruent AV stimuli (see Fig. 1 for a violin plot of the distribution of response times for each trial type). These findings suggest that the processes involved in combining incongruent auditory and visual information into a unified fused percept do not take any longer than those involved in combining congruent auditory and visual information, at least for the verbal, open-set task used here.

Congruent versus all McGurk analysis Although response times to congruent trials did not differ from those to McGurk fusion trials, it is possible that responses to congruent trials may differ from those to all McGurk trials when we collapse across fusion and non-fusion response types, as is often done in research using McGurk stimuli (e.g., Nahorna, Berthommier, & Schwartz, 2012; Nahorna et al., 2015). As in our previous analysis, we built a full model with stimulus type as a fixed effect, but here we collapsed across fusion and non-fusion responses – that is, trials were coded as either "congruent" or "McGurk." The reduced model was identical to the full model but lacked stimulus type as a fixed effect. A likelihood ratio test again indicated that response times did not differ by condition ($\chi^2_1 = 0.94$; $p = 0.33$). These findings contradict previous research showing that individuals respond to McGurk stimuli more slowly than congruent AV stimuli (Beauchamp et al., 2010; Nahorna et al., 2012; Tiippana et al., 2011), and instead suggest that at least in conditions in which participants are allowed to verbally respond with what they perceived rather than making a decision via a button box, response times to congruent syllables do not differ from those to McGurk syllables.

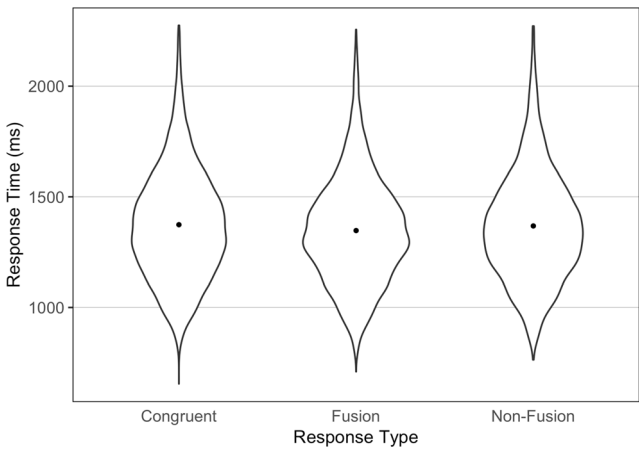

Fig. 1 Distribution of response times for each of the three trial types in Experiment 1. Dots represent the condition means

McGurk fusion versus McGurk non-fusion analysis Finally, given the scarcity of research comparing response times to McGurk fusion versus non-fusion responses, we sought to determine whether these response types differ in the speed with which they are responded to. To test this effect, we built a full model with response type (fusion versus non-fusion) as a fixed effect, and compared it to a reduced model lacking the fixed effect. Results of a likelihood ratio test indicated that response times to McGurk stimuli that resulted in a fused percept were faster than those that did not result in a fused percept ($\chi^2_1 = 4.69$; $p = 0.03$). Examination of the summary output of the full model indicated that response times were an estimated 14 ms faster for fused responses relative to non-fused responses ($\beta = -14.27$, $SE = 6.52$, $t = -2.19$, $p = 0.03$). Note that though significant at an α-level of .05, the magnitude of this effect is quite small ($d = 0.06$).

In Experiment 2, we assessed whether processing McGurk syllables requires more attentional resources than processing congruent syllables, despite the statistically indistinguishable verbal response times reported in this experiment.

Experiment 2

Method

Participants To attain the approved sample size of 85 participants, we collected data from 97 participants (ages 18–30 years) from the Carleton College community. Data collection occurred at different sites for Experiment 1 and Experiment 2, which ensured that no participant completed both studies. Five participants were excluded prior to performing any analyses due to technical difficulties or unintelligible responses. One participant was excluded for having poor accuracy at the secondary vibrotactile task and two participants were excluded for having mean fusion rates below chance levels. We only

analyzed data from the first 85 usable data files, after eliminating participants for the reasons described above. All procedures were approved by the Carleton College Institutional Review Board.

Stimuli The speech stimuli were identical to those in Experiment 1 and procedures were kept as similar as possible across the two experiments. However, in Experiment 2, participant responses did not trigger presentation of the blank ISI screen. This step was included in Experiment 1 because we wanted to indicate to participants that their verbal responses had been registered, but if this had occurred in Experiment 2, responses to vibrotactile stimuli would have initiated presentation of the blank screen, and if participants responded quickly enough, this would interfere with presentation of the speech stimuli. As in Experiment 1, the video file began approximately 275 ms before speech onset, but in this experiment, the video ended approximately 275 ms after speech offset, at which point a blank screen appeared while participants responded to the vibrotactile task and repeated the syllable they perceived.

Vibrotactile stimulation was presented via a custom-made apparatus attached to the index finger of each participant's non-dominant hand. The apparatus consisted of a 3D printed finger rest that contained a DC vibrating motor (akin to the vibrating mechanism in a cell phone), controlled via the digital output of a Cedrus RB-740 buttonbox. During each trial, participants were presented with short (100 ms), medium (150 ms), or long (250 ms) pulses from the vibrotactile stimulator. These pulses were shorter than those used in previous studies (e.g., 250 ms and 500 ms in Fraser et al., 2010; Gosselin & Gagné, 2011a, 2011b) because we wanted to make the secondary task more difficult. Each pulse length occurred an equal number of times. Participants were instructed to respond with their dominant hand after each vibrotactile stimulus by pressing one of three buttons on the button box corresponding to the stimulus they perceived. The button corresponding to "short" was always on the left, the "medium" button was in the middle, and the "long" button was on the right. The onset of the pulse varied from 60 ms before the onset of the auditory track (to ensure that even the shortest pulses overlapped with the speech to some extent) to 140 ms after auditory onset, in 50-ms increments. Response times were recorded from the onset of the vibrotactile pulse.

Procedure All participants completed three blocks of trials – an initial block to measure McGurk susceptibility, a block with the vibrotactile task alone to familiarize participants with the short, medium, and long pulses, and a main experimental block to measure the attentional costs associated with integrating congruent versus incongruent speech. The first block consisted of presentation of a randomly intermixed set of congruent and incongruent syllables without the secondary task. In that block, each incongruent stimulus token (six per stimulus) was presented twice in a randomized order, for a total of 48 incongruent trials (4 stimuli * 6 tokens * 2 repetitions), and an additional two tokens of each of the 11 congruent syllables served as fillers, for a total of 70 trials. Prior to beginning the task, participants completed three practice trials (two McGurk and one congruent). Responses collected in this block were transcribed offline to calculate the mean McGurk fusion rate for each participant, and we excluded data from participants who reported perceiving the McGurk effect below chance levels (assuming approximately 15 possible consonant responses, responding at chance levels corresponds to a fusion rate of 6.67%, so we eliminated participants who responded at rates below 7%). Given that simultaneously performing a secondary task reduces McGurk susceptibility (Alsius et al., 2005, 2007), this block was included solely for the purpose of excluding participants with extremely low fusion rates.

During the second block, participants were first presented with two successive pulses of the same duration in ascending order (i.e., two short pulses, followed by two medium pulses, followed by two long pulses). Before each pair, participants were informed about how they should classify the two pulses. After this brief exposure phase, participants were presented with 18 randomly intermixed trials (six each of short, medium, and long), and were asked to classify the pulses as short, medium, or long. If their accuracy during this practice block was worse than 75% (i.e., fewer than 14/18 correct), the block was repeated from the beginning (including the exposure phase). This block was included to ensure that participants were comfortable with classifying the pulses according to their duration before proceeding to the dual-task paradigm. For the second and third blocks, the vibrotactile stimulator was placed on a sound-absorbing pad under the testing desk to dampen the sound of the vibrations and reduce visual distraction. Although the vibrations may still have been audible, given that the congruent and incongruent trials contain the same types of vibrotactile stimuli, and congruence is manipulated within-subjects, audibility of the vibrotactile task should not systematically affect results.

After completing the vibrotactile familiarization block, participants received instructions on how to perform the tasks for the main experimental block, and completed eight practice trials with intermixed congruent and incongruent speech tokens and vibrotactile pulses. Within this final block, speech stimuli and vibrotactile pulses were presented in a randomized order, and speech congruity was intermixed. Participants were instructed to repeat aloud what they perceived while simultaneously performing the vibrotactile task. Participants were told that the speech task was most important and to focus attention on that task (Bourland-Hicks & Tharpe, 2002; Desjardins & Doherty, 2013; Downs, 1982; Fraser et al., 2010), but they should also attempt to perform the vibrotactile task to the best of their ability. The congruent and incongruent speech stimuli in this experiment were exactly the same as in Experiment 1, so each

participant completed 288 trials (144 congruent trials and 144 congruent trials). Participants had 2,000 ms to respond to the vibrotactile stimulus, and then there was a 500-ms ISI with a blank screen before the next trial.

Note that we chose not to include a baseline condition in which participants completed the vibrotactile task alone because the question of interest is not whether AV integration requires attention; rather, we are testing whether the attentional costs associated with integrating congruent and incongruent speech differ, so baseline performance is not necessary to answer this question.

Results and discussion

As in Experiment 1, we excluded trials with response times longer than 2,000 ms from further analysis (less than 1% of the data). Trials with extreme response times, as identified by our MAD exclusion criterion, were also excluded (1.58% of trials). Finally, we removed trials during which participants incorrectly classified vibrotactile stimuli as short, medium, or long (22.67% of the data). The reported analyses consisted of 9,366 congruent trials, 5,265 fusion responses, and 3,900 non-fusion responses. By-participant McGurk fusion rates ranged from 10.49% to 87.50%, and the mean fusion rate was 57.16%. The mean by-stimulus fusion rate ranged from 41.73% to 87.40%.

Unless otherwise specified, the analyses reported here follow those of Experiment 1, with the exception that response times in these analyses correspond to the secondary vibrotactile task, but in Experiment 1 they were in response to the speech itself. The first set of analyses compared congruent trials to McGurk fusion trials, and the second compared McGurk fusion to McGurk non-fusion responses. In Experiment 1, we analyzed differences between congruent and McGurk trials collapsed across response types to follow the convention of prior work. We did not conduct a parallel analysis for Experiment 2 because our focus was on assessing differences in the attentional costs associated with processing congruent stimuli and McGurk trials that resulted in fusions.

Congruent versus McGurk fusion analysis We first built a full model with stimulus type (congruent or McGurk fusion) as a fixed effect and compared this model to a reduced model that lacked the fixed effect for stimulus type (but was identical to the full model in all other respects) via a likelihood ratio test, which indicated that the effect of stimulus type was not significant ($\chi^2_1 = 0.48$; $p = 0.49$). Consistent with the results of Experiment 1, we did not find evidence for our hypothesis that response times to the vibrotactile task would be slower when processing McGurk fusions than congruent stimuli (see Fig. 2 for a violin plot of the distribution of response times for each trial type). These results suggest that the attentional resources required to process congruent stimuli and McGurk fusions do not differ, at least in the circumstances of this study.

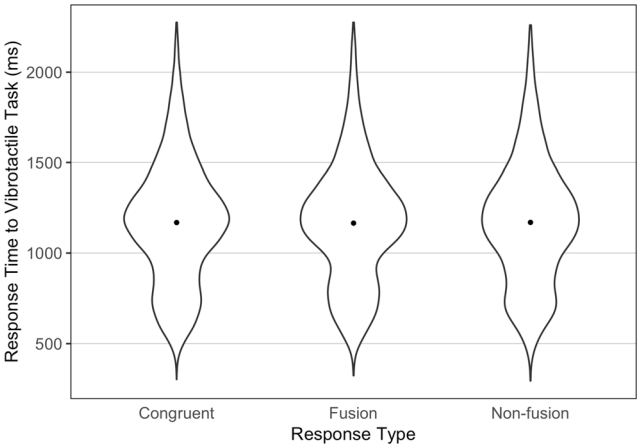

Fig. 2 Distribution of response times to the vibrotactile task for each of the three trial types in Experiment 2. Dots represent the condition means.

McGurk fusion versus McGurk non-fusion analysis Next, we compared response times to the vibrotactile task during McGurk trials that elicited a fused response to response times during those that did not. The full model contained McGurk response type (fusion or non-fusion) as a fixed effect, and the reduced model was identical to this model but lacked the fixed effect for response type. A likelihood ratio test indicated that there is no evidence that response times for McGurk fusion and non-fusion responses differ ($\chi^2_1 = 0.19$; $p = 0.66$). Taken with the results from the other two analyses in this experiment, these findings suggest that in the conditions assessed here, the amount of effort required to process and respond to congruent, McGurk fusion, and McGurk non-fusion syllables is indistinguishable (see Fig. 2).

Exploratory analysis comparing fusion rates in single- versus dual-task blocks Prior research has shown that McGurk fusion rates are lower when participants complete a simultaneous auditory (Alsius et al., 2005), visual (Tiippana et al., 2004), or tactile (Alsius et al., 2007) distractor task than in blocks without a dual-task. To assess whether we observed that finding in our data as well, we conducted an exploratory analysis (not part of the registered analysis plan) in which we compared fusion rates in the first block (speech-only) and the third block (speech plus vibrotactile task). Fusion rates were slightly higher in the speech-only block (M = 60.88) than in the speech plus dual-task block (M = 57.15), in line with the previous results – indeed, a model predicting fusion rates during McGurk trials using block as a fixed effect provided a significantly better fit for the data than a model without block ($\chi^2_1 = 4.22$; $p = 0.04$). The effect in this study is much smaller than has been reported previously, which may be a function of methodological differences. Our study differed from prior work in several key ways, notably that we used syllables rather than words (Alsius et al., 2005, 2007), used stimuli that were piloted to ensure that they regularly elicited fusion responses rather than factorial combinations of phonemes

(Tiippana et al., 2004), and we explicitly defined fusion responses and did not count non-auditory (Tiippana et al., 2004) or visual responses (Alsius et al., 2005, 2007) as fusions. Although it is conceivable that the difference between the speech-only block and the dual-task block is a function of the order in which the blocks were presented, an examination of the fusion rates over the course of the speech-only block did not reveal any systematic differences in fusion rates across time. It is therefore unlikely that this finding is attributable to order effects.

General discussion

In two experiments, we compared the time and effort necessary to process congruent AV speech, McGurk tokens that resulted in fusions, and McGurk tokens that did not result in fusions. Experiment 1 showed that the amount of time to initiate a verbal response to congruent and McGurk syllables was equivalent, but within the McGurk stimuli, those that resulted in fused responses were moderately faster than those that resulted in non-fused responses. Experiment 2 showed that response times to an unrelated vibrotactile task were identical for all stimulus types, suggesting that they incur similar processing costs.

Previous research has shown that response times to McGurk stimuli, typically collapsed across response types (fusion vs. non-fusion), are slower than those to congruent AV stimuli (Beauchamp et al., 2010; Nahorna et al., 2012; Tiippana et al., 2011). It is therefore somewhat surprising that we did not find any evidence for differences in response times to congruent and incongruent stimuli in Experiment 1. However, as discussed in the *Introduction*, task demands may have a substantial influence on response times to incongruent AV speech. The previous studies showing that McGurk stimuli are responded to more slowly than congruent stimuli consistently used closed-set tasks. These tasks require that participants indicate their responses via a button or key press, which limits responses to a small number of options (e.g., "b," "g," "d") and forces participants to categorize a percept that may be a poor exemplar of any phoneme category (Brancazio, 2004; Brancazio & Miller, 2005; Gentilucci & Cattaneo, 2005; Massaro & Ferguson, 1993; Rosenblum & Saldaña, 1992). As a result, participants may struggle to match the poor exemplar (e.g., A_bV_g that results in the perception of /d/) to a discrete, predefined category, thus requiring additional processing time. In contrast, congruent AV tokens (A_dV_d), which are likely to be very good exemplars of a given category, may be categorized more quickly.

The selective slowing for incongruent stimuli is less likely to occur in open-set tasks like the one we employed because participants are free to respond with whatever they perceived. In cases where a percept does not fit neatly into any category, participants can simply repeat an ambiguous syllable that falls somewhere between /b/ and /d/. Indeed, the research assistants who scored the participant responses informed us that some of the spoken syllables were difficult for them to categorize. Taken together, these results suggest that the method by which participants respond to syllables (forced-choice vs. open set) may affect study outcomes (see Basu Mallick et al., 2015; Clopper & Pisoni, 2007; Clopper et al., 2006; Colin et al., 2005, for examples of situations in which closed- and open-set tasks yield different results).

There are numerous reasons that the time required to process fusion and non-fusion responses might differ (see *Introduction*), but because several compelling arguments could be made in favor of slower response times to either fusion or non-fusion responses, we did not specify a directional hypothesis for this analysis. Thus, given the paucity of research on the topic, we assessed whether these two response types are processed at different rates, and demonstrated that response times to McGurk fusions were moderately faster than those to non-fusions. One outstanding question in the literature is why the same McGurk token tends to elicit fused responses in some participants but non-fused responses in others. It might be argued that participants only report non-fusion responses when they have failed to extract or integrate visual information on those trials. However, the fact that we showed that non-fused responses are actually longer than fused responses implies that those trials are not simply instances in which participants ignored visual input and reported what they heard. Instead, the results suggest that during these trials, participants attempted to integrate information from the auditory and visual modalities but this process broke down (or participants were distracted by the conflicting information) and integration was not achieved, resulting in increased processing time.

An implication of the difference in the time required to respond to trials that result in fusions and non-fusions is that researchers should use caution when collapsing across those response types. Prior work showing differences in the time required to respond to McGurk and congruent trials have typically included all McGurk trials in the analysis, regardless of whether they resulted in fusions. The results of Experiment 1 suggest that this may mask potential differences in the time and resources associated with the processing of these three types of AV stimuli. It is also worth noting that McGurk fusion and non-fusion trials differed in the speed with which they were responded to (Experiment 1) but not the effort necessary to process them (Experiment 2). These results also suggest caution in drawing conclusions about the attentional costs of a task from the time required to complete it.

Given that incongruent speech activates both speech-specific and general conflict-processing brain areas to a greater extent than congruent AV speech (Morís Fernández et al., 2017), and the McGurk effect is reduced when attention is divided (Alsius et al., 2005, 2007), we had expected that

processing incongruent stimuli would require more attentional resources than processing congruent AV speech. However, the results of the dual-task paradigm in Experiment 2 suggest that the attentional resources required to process AV speech do not differ for congruent and incongruent stimuli. Given the disparity between the results of Experiment 1 and prior work showing response-time differences for congruent and incongruent stimuli (e.g., Beauchamp et al., 2010; Massaro & Cohen, 1983; Tiippana et al., 2011), it may be that the null effect of stimulus type on effort is a function of the task used. That is, combining auditory and visual stimuli into a unified percept (as required by an open-set task) may be a rapid, automatic process whether the stimuli are congruent or incongruent, but *categorizing* the percept (as required by a closed-set task) may require more time or effort for incongruent than congruent stimuli. Note, however, that we did not include a closed-set task in either experiment, so this explanation is speculative. Thus, future work should assess whether the results reported here extend to a closed-set task. This could shed light on whether the process of categorization adds additional costs above and beyond those that may be incurred by combining auditory and visual speech information into a unified percept.

Despite the prevalence of the McGurk effect in speech perception research, there is converging evidence that processing McGurk stimuli differs in fundamental ways from processing congruent speech (Erickson et al., 2014; Morís Fernández et al., 2017; Van Engen et al., 2017). We show that, at least for the open-set speech task used here, congruent and McGurk stimuli do not differ in the speed or effort required to process them, but McGurk fusions are processed more quickly than non-fusions. Thus, researchers interested in studying the McGurk effect should distinguish between fusion and non-fusion responses, as collapsing across response types may obscure the interpretation of results of McGurk experiments. Further, the lack of a difference in response times between congruent and all McGurk stimuli in the open-set task we employed here, despite a substantial body of research showing slower responses for McGurk stimuli using closed-set tasks (Massaro & Cohen, 1983), suggests that task demands may play a crucial role in the McGurk effect (see also Basu Mallick et al., 2015; Colin et al., 2005), and future experiments should acknowledge this possibility.

Acknowledgements The authors thank Kristin Van Engen for helpful feedback on an earlier draft of the paper and the research assistants at Carleton College and Washington University in St. Louis who assisted with data collection and transcription. Carleton College supported this work.

Compliance with ethical standards

Open Practices Statement This Registered Report was approved in principle prior to data collection (see https://osf.io/8t7an/). All data, code, and stimuli are available at https://osf.io/z6kv3/.

References

Alsius, A., Möttönen, R., Sams, M. E., Soto-Faraco, S., & Tiippana, K. (2014). Effect of attentional load on audiovisual speech perception: Evidence from ERPs. *Frontiers in Psychology, 5,* 727.

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology: CB, 15*(9), 839–843.

Alsius, A., Navarra, J., & Soto-Faraco, S. (2007). Attention to touch weakens audiovisual speech integration. *Experimental Brain Research, 183*(3), 399–404.

Alsius, A., Paré, M., & Munhall, K. G. (2017). Forty Years After Hearing Lips and Seeing Voices: the McGurk Effect Revisited. *Multisensory Research, 31*(1-2), 111–144.

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

Basu Mallick, D., Magnotti, J. F., & Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review, 22*(5), 1299–1307.

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/

Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). fMRI-Guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 30*(7), 2414–2417.

Bourland-Hicks, C., & Tharpe, A. M. (2002). Listening effort and fatigue in school-age children with and without hearing loss. *Journal of Speech, Language, and Hearing Research: JSLHR, 45*(3), 573–584.

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology. Human Perception and Performance, 30*(3), 445–463.

Brancazio, L., & Miller, J. L. (2005). Use of visual information in speech perception: Evidence for a visual rate effect both with and without a McGurk effect. *Perception & Psychophysics, 67*(5), 759–769.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology: CB, 10*(11), 649–657.

Caramazza, A., & Brones, I. (1979). Lexical access in bilinguals. *Bulletin of the Psychonomic Society, 13*(4), 212–214.

Clopper, C. G., & Pisoni, D. B. (2007). Free classification of regional dialects of American English. *Journal of Phonetics, 35*(3), 421–438.

Clopper, C. G., Pisoni, D. B., & Tierney, A. T. (2006). Effects of open-set and closed-set task demands on spoken word recognition. *Journal of the American Academy of Audiology, 17*(5), 331–349.

Colin, C., Radeau, M., & Deltenre, P. (2005). Top-down and bottom-up modulation of audiovisual integration in speech. *The European Journal of Cognitive Psychology, 17*(4), 541–560.

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk–MacDonald effect: A phonetic representation within short-term memory. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology, 113*, 495–506.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*. https://doi.org/10.1037//0033-295x.82.6.407

Desjardins, J. L., & Doherty, K. A. (2013). Age-related changes in listening effort for various types of masker noises. *Ear and Hearing, 34*(3), 261–272.

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.

Erickson, L. C., Zielinski, B. A., Zielinski, J. E. V., Liu, G., Turkeltaub, P. E., Leaver, A. M., & Rauschecker, J. P. (2014). Distinct cortical locations for integration of audiovisual speech and the McGurk effect. *Frontiers in Psychology*, *5,* 534.

Forster, K. I., & Bednall, E. S. (1976). Terminating and exhaustive search in lexical access. *Memory & Cognition*, *4*(1), 53–61.

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.

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.

Gentilucci, M., & Cattaneo, L. (2005). Automatic audiovisual integration in speech perception. *Experimental Brain Research. Experimentelle Hirnforschung. Experimentation Cerebrale, 167*(1), 66–75.

Gilchrist, J. M., & Allen, P. M. (2015). Lexical decisions in adults with low and high susceptibility to pattern-related visual stress: A preliminary investigation. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00449

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

Gosselin, P. A., & Gagné, J.-P. (2011b). Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech Language and Hearing Research, 54*(3), 944–958.

Gottfried, J. A., & Dolan, R. J. (2003). The nose smells what the eye sees: Crossmodal visual facilitation of human olfactory perception. *Neuron, 39*, 375–386.

Grant, K. W., & Seitz, P. F. (1998). Measures of auditory–visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America, 104*(4), 2438–2450.

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.

Green, K. P., & Kuhl, P. K. (1991). Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology. Human Perception and Performance, 17*(1), 278–288.

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

Keane, B. P., Rosenthal, O., Chun, N. H., & Shams, L. (2010). Audiovisual integration in high functioning adults with autism. *Research in Autism Spectrum Disorders*, *4*(2), 276–289.

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

Lackner, J. R. (1977). Induction of illusory self-rotation and nystagmus by a rotating sound-field. *Aviation, Space, and Environmental Medicine, 48*(2), 129–131.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The Neighborhood Activation Model. *Ear and Hearing, 19*(1), 1–36.

Magnotti, J. F., Basu Mallick, D., Feng, G., Zhou, B., Zhou, W., & Beauchamp, M. S. (2015). Similar frequency of the McGurk effect in large samples of native Mandarin Chinese and American English speakers. *Experimental Brain Research. Experimentelle Hirnforschung. Experimentation Cerebrale, 233*(9), 2581–2586.

Massaro, D. W., & Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(5), 753–771.

Massaro, D. W., & Ferguson, E. L. (1993). Cognitive style and perception: the relationship between category width and speech perception, categorization, and discrimination. *The American Journal of Psychology, 106*(1), 25–49.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*. https://doi.org/10.1038/264746a0

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.

Morís Fernández, L., Macaluso, E., & Soto-Faraco, S. (2017). Audiovisual integration as conflict resolution: The conflict of the McGurk illusion. *Human Brain Mapping*. https://doi.org/10.1002/hbm.23758

Nahorna, O., Berthommier, F., & Schwartz, J.-L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *The Journal of the Acoustical Society of America, 132*(2), 1061–1077.

Nahorna, O., Berthommier, F., & Schwartz, J.-L. (2015). Audio-visual speech scene analysis: Characterization of the dynamics of unbinding and rebinding the McGurk effect. *The Journal of the Acoustical Society of America, 137*(1), 362–377.

Navarra, J., Alsius, A., Soto-Faraco, S., & Spence, C. (2010). Assessing the role of attention in the audiovisual integration of speech. *An International Journal on Information Fusion, 11*(1), 4–11.

Norrix, L. W., Plante, E., & Vance, R. (2006). Auditory-visual speech integration by adults with and without language-learning disabilities. *Journal of Communication Disorders, 39*(1), 22–36.

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin, 116*(2), 220–244.

R Core Team. (2016). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Retrieved from http://www.R-project.org/

Rips, L. J., & Shoben, E. J. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12:*1–20.

Rosenblum, L. D., & Saldaña, H. M. (1992). Discrimination tests of visually influenced syllables. *Perception & Psychophysics, 52*(4), 461–473.

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 487–494.

Saldaña, H. M., & Rosenblum, L. D. (1993). Visual influences on auditory pluck and bow judgments. *Perception & Psychophysics, 54*(3), 406–416.

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.

Shankar, M. U., Levitan, C. A., Prescott, J., & Spence, C. (2009). The influence of color and label information on flavor perception. *Chemosensory Perception*, *2*(2), 53–58.

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.

Soto-Faraco, S., & Alsius, A. (2007). Conscious access to the unisensory components of a cross-modal illusion. *Neuroreport, 18*(4), 347–350.

Soto-Faraco, S., Navarra, J., & Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition, 92*(3), B13–B23.

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.

Strand, J. F., Cooperman, A., Rowe, J., & Simenstad, A. (2014). Individual differences in susceptibility to the McGurk effect: Links with lipreading and detecting audiovisual incongruity. *Journal of Speech, Language, and Hearing Research: JSLHR, 57*(6), 2322–2331.

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.

Talsma, D., & Woldorff, M. G. (2005). Selective attention and multisensory integration: Multiple phases of effects on the evoked brain activity. *Journal of Cognitive Neuroscience, 17*(7), 1098–1114.

Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *The European Journal of Cognitive Psychology, 16*(3), 457–472.

Tiippana, K., Puharinen, H., Möttönen, R., & Sams, M. (2011). Sound location can influence audiovisual speech perception when spatial attention is manipulated. *Seeing and Perceiving, 24*(1), 67–90.

Toscano, J. C., & Allen, J. B. (2014). Across- and within-consonant errors for isolated syllables in noise. *Journal of Speech, Language, and Hearing Research: JSLHR, 57*(6), 2293–2307.

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition, 96*(1), B13–B22.

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 der Burg, E., Brederoo, S. G., Nieuwenstein, M. R., Theeuwes, J., & Olivers, C. N. L. (2010). Audiovisual semantic interference and attention: evidence from the attentional blink paradigm. *Acta Psychologica, 134*(2), 198–205.

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.

Van Engen, K. J., Xie, Z., & Chandrasekaran, B. (2017). Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. *Attention, Perception & Psychophysics, 79*(2), 396–403.

Zampini, M., & Spence, C. (2004). The role of auditory cues in modulating the perceived crispness and staleness of potato chips. *Journal of Sensory Studies, 19*(5), 347–363.