Attention, Perception, & Psychophysics (2026) 88:78
https://doi.org/10.3758/s13414-025-03206-3

The dual-task costs of audiovisual benefit: Effects of noise and "native" speaker status

Violet A. Brown[1] · Adina Holloway[1] · Amadou Touré[1] · Salma Ali[1] · Alyssa Alvarez[1] · Tiffany Nyamao[1] · Yuxin Lin[1] · Ostap Hrebeniuk[1] · Julia F. Strand[1]

Received: 29 May 2025 / Accepted: 12 November 2025

Abstract

Listeners typically understand speech more accurately when they can see and hear the talker relative to hearing alone. However, seeing the talker's face does not necessarily reduce the cognitive costs associated with processing speech as measured by dual-task costs. In difficult listening conditions, dual-task response times may be faster for audiovisual than audio-only speech, but when listening conditions are easy, the presence of a talking face may have no effect on dual task responses or even slow responses relative to listening alone. The current study expanded upon this work by including samples of both native and nonnative English speakers and assessing speech intelligibility, subjective listening effort (Experiment 1), and dual-task costs (Experiment 2) for audio-only and audiovisual speech across multiple noise levels. We found that seeing the talker reduces dual-task costs only in difficult listening conditions in which the visual information is necessary to accurately identify the speech. The effects of background noise and speech modality were robust within groups of native as well as nonnative listeners, suggesting that if researchers are interested in studying general phenomena related to speech processing (i.e., rather than specifically studying how language background affects results), these effects would have emerged regardless of whether the sample was limited to native speakers of English. However, the magnitude of some effects differed for native and nonnative listeners.

Keywords Audiovisual speech · Listening effort · Nonnative speech processing · Dual-task

Speech identification in noise is greatly improved when the listener can see the talker's face in addition to hearing their voice (often referred to as "audiovisual benefit"; Erber, 1969; Sumby & Pollack, 1954). These effects are particularly pronounced in adverse listening conditions, including when the speech is presented in higher levels of background noise (e.g., Brown & Strand, 2019; Ross et al., 2006; Sumby & Pollack, 1954) or when it is acoustically degraded via vocoding (Blackburn et al., 2019). Audiovisual benefit occurs in part because the movements of a talking face provide phonetic information that is complementary to the auditory information and can therefore be used to disambiguate words that are acoustically similar but visually distinct (e.g., "lap" and "lag"; Grant et al., 1998; Grant & Walden, 1996). In other words, visual cues may help reduce lexical competition, particularly when the listening conditions lead to a degraded acoustic input.

The reduction in lexical competition afforded by the visual signal has consequences beyond speech intelligibility. Indeed, there is mounting evidence that visual speech cues can reduce the cognitive costs associated with processing speech in noisy conditions, often referred to as "listening effort" (see Francis & Love, 2020, for a review). Although listening effort is a multifaceted construct that has been measured in numerous ways (e.g., Alhanbali et al., 2019; Strand et al., 2018, 2021), one of the most common measurement techniques involves implementing dual-task paradigms.[1] In the context of audiovisual speech processing,

✉ Violet A. Brown
violetbrown@carleton.edu

✉ Julia F. Strand
jstrand@carleton.edu

[1] Carleton College, One North College St, Northfield, MN 55057, USA

[1] Given the multidimensional nature of listening effort, the numerous ways to measure each aspect of the construct, and the lack of agreement about how best to define and measure the construct broadly and the aspects specifically, we will use the term "dual-task costs" rather than "listening effort" going forward (see also `Brown, 2025). However, we mention listening effort to draw attention to relevant literature and facilitate potential future meta-analyses.

Published online: 23 February 2026

participants are typically asked to listen to audio-only and audiovisual speech while completing a secondary response time task. Given the assumption that cognitive resources are finite (Kahneman, 1973), slowed responses on the secondary task are taken as an indication of greater cognitive recruitment for the speech task. Consistent with this claim, dual-task costs tend to be more pronounced in listening conditions that are presumed to be more difficult—like when the background noise is louder (e.g., Sarampalis et al., 2009; Seeman & Sims, 2015; Strand et al., 2018), when the speech presented in an unfamiliar accent (Brown et al., 2020), when reverberation is present (Picou et al., 2016), and in other adverse listening conditions (see Gagné et al., 2017, for a review).

Dual-task studies of audiovisual speech processing have shown that when the listening conditions are difficult (e.g., at low signal-to-noise ratios; SNRs) seeing the talker's face in addition to hearing their voice *decreases* dual-task costs relative to listening alone. However, when the listening conditions are very easy—like when the background noise is set to a relatively low level—seeing the talker's face either has no effect on dual-task costs relative to hearing alone, or may even *increase* costs (Brown, 2025; Brown & Strand, 2019). The explanation for these findings is that when the acoustic speech signal is degraded, the complementary phonetic cues provided by the talking face reduce the cognitively demanding process of lexical competition (Kuchinsky et al., 2013; Wagner et al., 2016). Thus, even if additional cognitive resources must be recruited to process the visual channel in addition to the auditory channel, the cognitive benefits outweigh any potential costs associated with processing information from two sensory channels. When the listening task is easy, however, the two-channel cost does not come with the benefit of reduced lexical competition. Thus, when the visual signal is not necessary for successful speech identification (i.e., in very easy listening conditions), there is either a net-neutral effect or an overall increase in dual-task costs in audiovisual relative to audio-only settings, depending on the overall difficulty of the task (Brown, 2025).

Taken together, there is clear consensus that seeing the talker facilitates speech identification in noise, and accumulating evidence that the dual-task costs associated with processing audiovisual speech differ depending on the difficulty of the speech task. However, nearly all the work supporting these conclusions has been conducted on participants who report being "native" speakers of English.[2]

Indeed, even beyond research specifically on *audiovisual* speech processing, research on spoken word identification and speech perception more generally tends to limit samples to native speakers (unless, of course, the research question specifically concerns native speaker status or bilingualism; Strand et al., 2024). Although this tendency may be driven in part by adherence to the conventions of the field, researchers may also limit their samples out of concern that the effects of interest may only appear in highly proficient speakers or in those who have been exposed to the language from birth.

Research on audiovisual spoken word identification has provided some evidence that nonnative listeners derive less audiovisual benefit for speech identification than native listeners, presumably because nonnative listeners have less experience mapping English phonemes to their corresponding mouth movements ("visemes"; Drijvers & Özyürek, 2020; Xie et al., 2014; Yang et al., 2024). Although language background appears to interact with presentation modality when the outcome measure involves speech *identification*, no work to date has assessed whether native and nonnative listeners also differ in the extent to which seeing the talker affects *dual-task costs*. In fact, research on differences in dual-task costs between native and nonnative listeners is quite limited, even for audio-only speech. The only study we have identified on the topic demonstrated that nonnative listeners exhibit slower response times to a secondary task than native listeners (Peng & Wang, 2019), suggesting that processing speech in one's nonnative language is more cognitively demanding than processing speech in one's native language (see also Borghini & Hazan, 2018).

This finding follows from previous work demonstrating that nonnative listeners tend to show poorer speech identification accuracy relative to native listeners (Black & Hast, 1962; Scharenborg & van Os, 2019), and tend to be more adversely affected by both background noise (Cooke et al., 2008) and lexical difficulty (Bradlow & Pisoni, 1999; Strand et al., 2024). These findings have typically been attributed to limited experience with the nonnative language as well as "nonselective lexical access," whereby lexical items from multiple languages are activated in multilingual individuals (Hintz et al., 2023). That is, if speech identification is more difficult for LX listeners, and adverse listening conditions tend to increase dual-task costs (e.g., Gagné et al., 2017), processing costs should be greater for nonnative relative to native listeners. However, it is unclear whether the effect of seeing the talker on dual-task costs differs by language background.

On the one hand, less audiovisual *intelligibility* benefit for nonnative relative to native listeners (Xie et al., 2014) may imply less audiovisual benefit at the level of cognitive effort as well. Indeed, if the visual signal eliminates lexical competitors more robustly for native than nonnative listeners, the

[2] There are limitations to using the terms "native" and "nonnative" to describe language groups, but when describing previous work in this area, we will use these terms for consistency (see `Cheng et al., 2021; `Strand et al., 2024; `Vulchanova et al., 2022). In our own work, however, we prefer to use the terms L1 ("native") and LX ("nonnative"; see `Dewaele, 2018; Strand et al., 2024).

additional competition experienced by nonnative listeners should not only impair intelligibility, but also slow response times to secondary tasks (i.e., increase effort). On the other hand, nonnative listeners may show larger benefits from the visual signal (as measured using the dual-task paradigm) because identifying speech in one's nonnative language is a cognitively demanding task (Borghini & Hazan, 2018; Peng & Wang, 2019) and audiovisual effort benefits tend to be more pronounced in difficult listening conditions (Brown, 2025). The current study attempts to adjudicate between these competing predictions.

The current study

Given limited research on the independent and joint effects of language background and audiovisual speech on dual-task costs, the current study assesses how seeing the talker affects dual-task costs at multiple levels of background noise, and whether these effects are moderated by the language background of the participants. In addition to addressing theoretical questions that will illuminate the cognitive mechanisms underlying nonnative speech processing (see above), including more representative samples rather than limiting our sample to only native English speakers will enhance the generalizability of these results while simultaneously counteracting existing inequities in the field (see Cheng et al., 2021; Strand et al., 2024). Finding qualitatively similar results for both language groups would supplement the argument we have made previously that the common practice of excluding nonnative listeners from our samples is often unnecessary *when the goal of the research is to establish general findings that should apply to most listeners* (e.g., effects of seeing the talker, background noise; Strand et al., 2024). Of course, the extent to which limiting the sample to native speakers affects study outcomes will vary from one study to the next. . But at the very least, obtaining qualitatively similar results would highlight another research area for which patterns of results are likely to persist regardless of the language backgrounds of the participants. These results would suggest that research on the dual-task costs associated with speech processing might benefit greatly from increased generalizability with little to no cost.

The primary goal of this study is to address theoretical questions described above regarding audiovisual speech processing and dual-task costs for native and nonnative listeners (Experiment 2). However, research on differences in audiovisual intelligibility benefit between language groups is rather limited, so we will also assess speech intelligibility in audio-only and audiovisual conditions across noise levels for the two language groups. Given that dual-tasking may affect performance on the speech identification task relative to identifying speech in isolation (e.g., Fraser et al., 2010), we will run a version of the experiment without a secondary task to obtain a clean measure of speech intelligibility (Experiment 1). This series of experiments will be conducted online with samples from the Prolific population (Palan & Schitter, 2018), which includes individuals with more diverse demographics than most studies conducted on college campuses, including variables such as age, education level, country of residence, and language background (e.g., see demographic details in the Supplementary Materials associated with Strand et al., 2024). Replicating previous work on differences in audiovisual benefit between native and nonnative listeners (Xie et al., 2014) will provide insight about the generalizability of the findings or the constraints on generalizability.

Listening effort is multidimensional, and different measures tap into different features of the underlying construct (Alhanbali et al., 2019; Francis & Love, 2020; McGarrigle et al., 2014; Strand et al., 2018, 2021). Indeed, dual-task costs reveal the online behavioral consequences of effortful listening; pupillometry (Beatty, 1982; Kramer et al., 2012), heart rate variability (Seeman & Sims, 2015), and skin conductance (Mackersie & Cones, 2011) reveal the physiological toll effortful listening has on the listener in the moment; recall shows downstream consequences of effortful listening (Brown & Strand, 2019; Rabbitt, 1968; Sommers & Phelps, 2016); and subjective self-reports of effort highlight the listener's experience during the task (Herrmann & Johnsrude, 2020). We opted to measure dual-task costs because we are interested in the cognitive mechanisms associated with combining auditory and visual speech signals, and dual-task paradigms give information about cognitive costs in real time (see Gagné et al., 2017; Kuchinsky et al., 2024 for reviews of the use of dual-task measures in speech research). Given the large number of tasks used to measure listening effort and the multidimensional nature of the construct, it is unsurprising that different tasks purporting to measure "listening effort" for audiovisual speech show fundamentally different patterns of results (see Brown & Strand, 2019). Thus, in addition to assessing speech intelligibility, Experiment 1 will measure subjective self-reported listening effort using a questionnaire that is widely used in the literature: The NASA Task Load Index (NASA-TLX; Hart & Staveland, 1988).

Experiment 1 (intelligibility and subjective self-report) and Experiment 2 (dual-task costs) will use different samples of online participants to shed light on audiovisual benefit—or perhaps detriment—across a range of noise levels for native and nonnative speakers of English. Rather than using the terms "native" and "nonnative," we instead adopt the terms "L1" ("native") and "LX" ("nonnative") to distinguish between individuals who learned English first (L1) and those who learned it after another language (see Cheng et al., 2021; Strand et al., 2024).

Experiment 1

Method

Data, materials, and code for Experiment 1 are available online at https://osf.io/m29rw/. The preregistration document, including details regarding sample size justification for each experiment, is available at https://osf.io/ctmdb.

Participants

To reach our preregistered sample size of 102 participants per group, we recruited 109 L1 participants and 107 LX participants online via Prolific (www.prolific.co). Participants were limited to Prolific users who were between the ages of 18 and 45 and had self-reported normal hearing and normal or corrected-to-normal vision. We limited L1 participants to Prolific users with American IP addresses[3] who had previously specified that English was their first language. LX participants were those who reported any language other than English as their first language when they registered with Prolific; this group was not geographically constrained. We replaced participants whose language background on Prolific did not match what they self-reported in our demographic questionnaire (L1: $N = 2$, LX: $N = 3$), as well as those whose accuracy at the speech identification task was worse than three standard deviations below the mean accuracy for any noise-by-modality condition, but only if the accuracy was also below 90% (L1: $N = 5$; LX: $N = 2$). Participants were compensated at a rate of $7 for 35 min of participation. Carleton College's Institutional Review Board approved all research procedures. Data collection occurred between March 7 and April 11, 2024, and on September 30 and October 1, 2024.

Participants were 19–45 years old (L1: $M = 32.3$, $SD = 7.01$; LX: $M = 28.4$, $SD = 5.5$), and LX participants learned English at an average age of 7.6 years ($SD = 2.9$) and reported relatively high levels of self-reported proficiency at understanding, reading, speaking, and writing English (means ranging from 7.57–9.2 on a 10-point scale) In all cases self-reported proficiency was lower for the LX group than the L1 group. For the majority of LX listeners, English was the second language learned and the second most dominant language.

[3] We opted to limit our sample to users with American IP addresses to more closely match samples that have historically been studied in the speech literature. We did not exclude listeners who spoke a dialect of English other than American English.

Stimuli

Words Speech stimuli consisted of 300 words (50 per modality-by-SNR condition) from the SUBTLEX-US database (Brysbaert et al., 2012). We excluded articles, conjunctions, rare words (log-frequencies less than three), and words with more than two syllables or five phonemes. The words were randomly divided into six lists and each list was used in each of the conditions approximately the same number of times across participants. Within each list, 70% of the words were predominantly classified as nouns (Brysbaert et al., 2012), and the sublists for each condition maintain this 70% noun composition.[4] Participants were randomly assigned to six groups that varied the assignment of word lists to conditions. Thus, although each participant only heard each word once, across participants every word appeared in every condition. Speech stimuli were recorded by an L1 English speaker without a strong regional accent at 16-bit, 44100 Hz using a Shure KSM-32 microphone with a plosive screen, and visual stimuli were recorded with a Panasonic AG-AC90 camera (these stimuli come from the same set used by Brown, 2025). We edited speech stimuli and equated for root-mean-square amplitude using Adobe Audition. Speech files were leveled to -24 dB.

Noise Background noise consisted of steady-state speech-shaped noise that matched the long-term average spectrum of the word stimuli (Winn, 2018). Given that the current studies use isolated words as targets, we opted to use steady-state noise rather than an informational masker (e.g., two-talker babble) to help ensure that participants could distinguish between the target talker and the masker. The level of the speech was held constant and the noise was adjusted to create signal-to-noise ratios (SNRs) of 10 (easy condition), −2 (medium condition), and −8 (hard condition). SNRs were selected via in-lab pilot testing to produce audio-only identification accuracies in the ~40–50% range in the hardest noise level and near ceiling accuracy in the easy condition.

[4] We controlled the proportion of words that could be used as nouns because we had originally planned to use a dual-task paradigm that involved making speeded noun judgments (`Picou & Ricketts, 2014; `Strand et al., 2018). However, while running Experiment 1, we realized that L1 and LX listeners may differ in their baseline performance on a noun judgment task (i.e., regardless of listening effort), so we switched to a nonverbal dual-task paradigm. As a result, the preregistration for Experiment 1 linked above contains a description of the noun judgment task for Experiment 2 which we did not use. An updated preregistration for Experiment 2 detailing the tone classification task we actually used is linked to in the Experiment 2 Methods section.

Procedure

The experiment was created and conducted via Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). Participants were instructed to wear headphones and complete the task in a quiet space. Before beginning, participants completed a sound check to ensure they could comfortably hear the audio and were asked not to adjust the volume after this initial phase. Next, participants completed a headphone screening consisting of six trials containing three 200 Hz sinusoidal tones, and participants were asked to identify which tone was the quietest (a task that cannot be reliably completed with speakers). Participants could only continue if they correctly identified the quietest tone in all six trials of the screening. If they failed on the first attempt, they had an opportunity to repeat the task one additional time (see Woods et al., 2017 for additional stimulus details).

After the headphone screening, participants completed the main word identification task, which involved identifying words in audio-only and audiovisual conditions at three SNRs (−8, 2, and +10). Each trial consisted of an isolated word in one of six SNR-by-modality conditions. After each word was presented, participants typed what they heard in a text box, guessing when unsure. Participants pressed the enter key to move on, and the next trial began after an interstimulus interval randomly selected from 1,500 ms to 3,000 ms, during which a fixation cross appeared on the screen. The word task was divided into six blocks of 50 words each, with each block consisting of a single SNR-by-modality condition. Participants were randomly assigned to one of six list assignments, and the order of the blocks within a list assignment was randomized. Each participant heard each word exactly once, but each word appeared in every condition approximately the same number of times across participants. Participants completed six practice trials (using stimuli that did not appear in the main task) before beginning the main task, including one word in each of the noise-by-modality conditions. Participants did not receive feedback, and we did not evaluate performance on the practice trials.

After every 10 words within a single condition, participants completed the NASA-TLX to measure subjective listening effort (Hart & Staveland, 1988). This questionnaire is commonly used in the listening effort literature, and asks participants to rate subjective effort, performance, frustration, and mental demand.[5] Participants provided responses via an unnumbered sliding scale; the position of the slider on the unnumbered scale was transformed to a 1–100 scale for analysis. The questions were listed in a consistent order (mental demand, performance, effort, and frustration) following each block of words.

Finally, after the main task, participants completed a demographic and language background questionnaire that was adapted from the Language Proficiency and Experiment Questionnaire (Marian et al., 2007; see https://osf.io/m29rw/ for complete questionnaire).

Results and discussion

Accuracy at identifying the speech was scored on a trial-by-trial basis such that correct responses were scored as 1 and incorrect responses were scored as 0. Prior to scoring trials for accuracy, we removed extraneous punctuation and modified responses that were homophones of the correct response (e.g., if the target word was "heard," participants received credit for typing the word "herd"). To correct misspellings in a systematic way that allows phonologically probable misspellings, we entered all incorrect responses into a software called "Ponto" (Kessler, 2017; see Strand et al., 2024, for details of implementation in similar research). This system flexibly scores participant responses by assigning each answer a value that represents the phonological similarity to the target word. Given that LX listeners typically have less experience with English spelling, this flexible scoring method serves as a systematic way to tease apart misspellings and mishearings. For example, for the target word "castle," the responses "cassel," "castel," and "casttle" were all counted as correct, given that they are phonologically plausible misspellings of the target. The flexible scoring method (Ponto; Kessler, 2017) corrected 4.6% of previously incorrect trials (see R script for details).

Given the binary nature of the outcome (correct/incorrect), intelligibility data were analyzed using logistic regression assuming a binomial data-generating process with a logit link function. Subjective effort data from the NASA-TLX questionnaire were analyzed assuming a Gaussian data-generating process and an identity link function. All data were analyzed using linear mixed effects models via the *lme4* (Bates et al., 2014) package in R (Version 4.2.2; R Core Team, (R 2022)). Statistical significance was evaluated by comparing nested models differing only in the effects of interest via likelihood ratio tests, and we provide coefficient estimates for effects. All reported coefficients were derived from models lacking higher-order effects to avoid interpreting lower-order terms in the presence of an interaction. Specifically, when testing main effects, we controlled for all other variables. When testing two-way interactions, we included all lower-order terms and controlled for all other variables, but only included the two-way interaction of interest.

We first analyzed the L1 and LX participants separately and then combined the data from the two groups to evaluate

[5] Note the full scale also includes questions on physical demand and temporal demand, which we omitted to omit due to a lack of relevance to the speech identification task and to avoid participant confusion.

the effects of language background on spoken word identification accuracy and subjective effort. For analyses conducted on L1 and LX participants separately, we attempted to model the maximal random effects structure justified by the design. For intelligibility analyses, this included random intercepts for participants and words, and by-participant as well as by-word random slopes for noise and modality. For subjective effort analyses, this included random intercepts for participants as well as by-participant random slopes for noise and modality, but no word-level random effects (because responses corresponded to blocks of words rather than individual words). Any deviations from this random effects structure are explicitly noted below. For intelligibility analyses of the combined L1/LX data, we opted not to attempt to model by-word random slopes for language group because we expected little variability in the effect of participant group across words (note that by-participant random slopes for language group are not justified because this is a between-subjects variable). Variables were dummy coded such that the easiest noise level, audio-only, and L1 were coded as 0 (where appropriate, we re-leveled models with the medium noise level as the reference level to obtain the medium/hard comparison).

Identification accuracy

Mean speech identification accuracies for both experiments are presented in Table 1. The results of the model comparisons are shown in Tables 2 and 3, and coefficient estimates for each model are reported below in the text. The models reported below are based on 30,495 observations for L1 participants and 30,588 observations for LX participants (a full dataset for each group would include 102 participants × 300 words = 30,600 observations; the missing items appear to be from participants experiencing technical issues during data collection).

L1 All models for analyzing intelligibility data from L1 listeners excluded by-noise random slopes for participants to facilitate model convergence. As would be expected,

Table 1 By-condition identification accuracy (percentage correct) in Experiments 1 and 2, and mean response times (ms) in Experiment 2 (grouped by participant such that standard deviations reflect variability across participant means)

SNR	Language group	Experiment 1 mean speech identification accuracy (*SD*)		Experiment 2 mean speech identification accuracy (*SD*)		Experiment 2 mean response time (*SD*)	
		AO	AV	AO	AV	AO	AV
10 (easy)	L1	96.94 (3.12)	97.66 (2.64)	95.33 (5.89)	95.92 (6.56)	1,456 (501)	1,451 (513)
	LX	89.78 (8.08)	89.90 (8.97)	89.63 (8.98)	91.04 (6.97)	1,579 (443)	1,559 (491)
−2 (medium)	L1	82.48 (8.16)	89.78 (5.81)	86.81 (10.24)	90.08 (9.54)	1,508 (549)	1,460 (503)
	LX	69.39 (12.91)	77.43 (12.56)	77.90 (11.63)	84.04 (9.08)	1,640 (514)	1,672 (498)
−8 (hard)	L1	49.00 (12.84)	72.05 (11.84)	68.61 (13.06)	80.66 (10.63)	1,634 (532)	1,567 (551)
	LX	38.74 (15.91)	54.88 (16.88)	56.23 (12.63)	70.49 (11.39)	1,846 (612)	1,745 (612)

Table 2 Results of model comparisons (likelihood ratio test; LRT) testing the effects of noise, modality, and their interaction on word identification accuracy and subjective listening effort for the L1 and LX groups separately (Experiment 1)

Effect tested	Larger model (italicized term is omitted in the reduced model)	Group	Identification accuracy LRT	Listening effort LRT
Main effect of noise	*noise* + modality	L1	$\chi^2_2 = 565.47, p < .001$	$\chi^2_2 = 132.34, p < .001$
		LX	$\chi^2_2 = 285.18, p < .001$	$\chi^2_2 = 136.51, p < .001$
Main effect of modality	noise + *modality*	L1	$\chi^2_1 = 132.61, p < .001$	$\chi^2_2 = 14.77, p < .001$
		LX	$\chi^2_1 = 84.72, p < .001$	$\chi^2_2 = 12.65, p < .001$
Interaction between noise and modality	noise + modality + *noise:modality*	L1	$\chi^2_2 = 104.77, p < .001$	$\chi^2_2 = 99.72, p < .001$
		LX	$\chi^2_2 = 102.77, p < .001$	$\chi^2_2 = 57.53, p < .001$

Table 3 Results of model comparisons (likelihood ratio test; LRT) for word identification accuracy and subjective listening effort data, L1 and LX data combined (Experiment 1)

Effect tested	Larger model (italicized term is omitted in the reduced model)	Identification accuracy LRT	Listening effort LRT
Main effect of language group	noise + modality + *participant_group*	$\chi^2_1 = 60.71$, $p < .001$	$\chi^2_1 = 23.92$, $p < .001$
Interaction between noise and language group	noise + modality + group + *noise:group*	$\chi^2_2 = 18.48$, $p < .001$	$\chi^2_2 = 4.99$, $p = .08$
Interaction between modality and language group	noise + modality + group + noise:modality + noise:group + *modality:group*	$\chi^2_1 = 11.11$, $p < .001$	$\chi^2_1 = 0.73$, $p = .39$
Interaction among noise, modality, and language group	noise + modality + group + noise:modality + noise:group + modality:group + *noise:modality:group*	$\chi^2_2 = 6.34$, $p = .04$	$\chi^2_2 = 15.91$, $p < .001$

identification accuracy was significantly affected by noise such that accuracy was higher in the easy than the medium SNR ($B_{medium} = -1.78$, $SE = 0.14$, $z = -13.07$, $p < .001$) and higher in the easy than the hard SNR ($B_{hard} = -3.97$, $SE = 0.14$, $z = -29.10$, $p < .001$), controlling for modality. Re-leveling the noise variable revealed that accuracy was significantly higher in the medium than the hard SNR ($B_{hard} = -2.19$, $SE = 0.08$, $z = -27.93$, $p < .001$). We also found a robust effect of modality, with higher identification accuracy for audiovisual relative to audio-only speech, controlling for noise ($B_{AV} = 1.13$, $SE = 0.08$, $z = 13.88$, $p < .001$).

The interaction between noise and modality was significant such that the audiovisual intelligibility benefit was larger in the hard than the medium SNR ($B_{AV*hard} = 0.78$, $SE = 0.10$, $z = 8.06$, $p < .001$) and larger in the medium than the easy SNR ($B_{AV*medium} = 0.55$, $SE = 0.16$, $z = 3.40$, $p < .001$). Unsurprisingly, therefore, the easy-hard comparison was also significant ($B_{AV*hard} = 1.33$, $SE = 0.16$, $z = 8.36$, $p < .001$). These results are consistent with previous research showing larger effects of seeing the talker on intelligibility at more difficult SNRs (Sumby & Pollack, 1954), presumably because the visual cues provided by the talking face have greater opportunity to benefit speech identification in louder levels of background noise (see Fig. 1).

LX The results for the LX participants were qualitatively the same as those for the L1 group. Controlling for modality, identification accuracy was higher in the easy than the

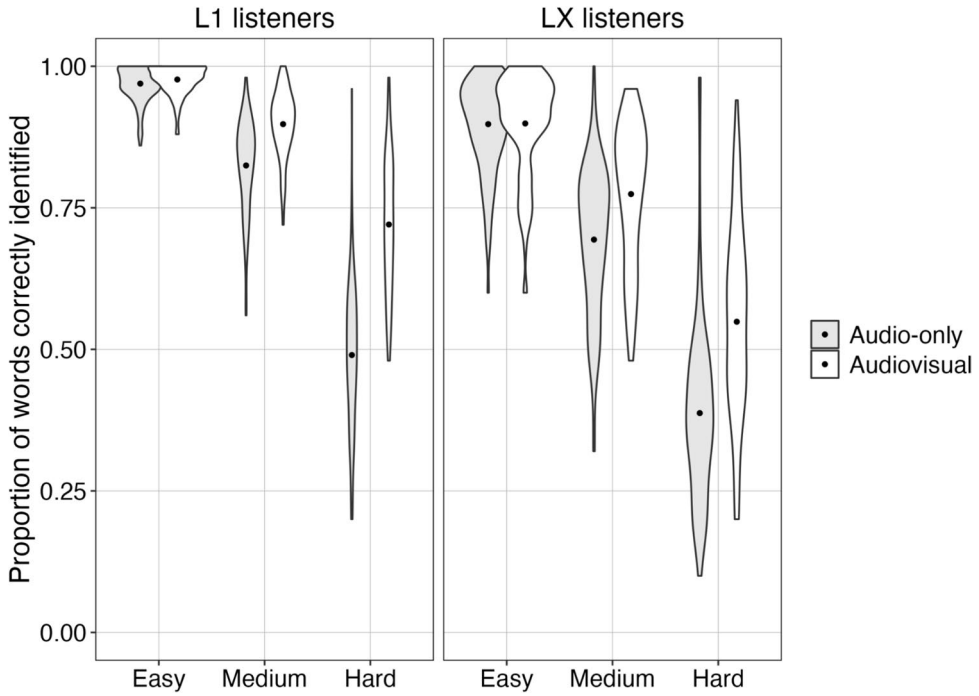

Fig. 1 Violin plot showing by-participant identification accuracies grouped by language group, noise, and modality. Dots indicate mean values for each condition

medium SNR ($B_{medium} = -1.72$, $SE = 0.09$, $z = -18.16$, $p < .001$), higher in the medium than the hard SNR ($B_{hard} = -1.70$, $SE = 0.07$, $z = -23.97$, $p < .001$), and higher in the easy than the hard SNR ($B_{hard} = -3.42$, $SE = 0.12$, $z = -27.73$, $p < .001$). Seeing the talker also improved speech identification accuracy overall, controlling for noise level ($B_{AV} = 0.67$, $SE = 0.06$, $z = 10.62$, $p < .001$).

There was again an interaction between noise and modality, and the pattern of results mirrored those described above for L1 listeners (Fig. 1). The effect of modality was stronger in the hard than the medium SNR ($B_{AV*hard} = 0.50$, $SE = 0.08$, $z = 6.39$, $p < .001$) and stronger in the medium than the easy SNR ($B_{AV*medium} = 0.52$, $SE = 0.10$, $z = 5.45$, $p < .001$). Consistent with the L1 data, the modality effect was also stronger in the hard than the easy SNR ($B_{AV*hard} = 1.02$, $SE = 0.10$, $z = 10.23$, $p < .001$).

Combined analysis Next, we combined the L1 and LX data to assess the effects of language background on speech intelligibility in different noise levels and modalities. Relative to L1 participants, LX participants had lower accuracy overall ($B = -0.98$, $SE = 0.12$, $z = -8.41$, $p < .001$) and showed smaller effects of modality ($B_{AV*group} = -0.28$, $SE = 0.08$, $z = -3.42$, $p < .001$). The interaction between participant group and noise level was also significant: Relative to easy noise, the negative effects of both medium ($B_{group*medium} = 0.28$, $SE = 0.11$, $z = 2.68$, $p = .007$) and hard ($B_{group*hard} = 0.54$, $SE = 0.13$, $z = 4.28$, $p < .001$) noise were slightly *less* pronounced for LX listeners than L1 listeners, and effects of hard relative to medium noise were also less pronounced for LX listeners ($B_{group*hard} = 0.26$, $SE = 0.08$, $z = 3.29$, $p = .001$).

This finding may at first glance appear somewhat puzzling and inconsistent with the condition means reported in Table 1. Indeed, the effects of noise on the raw (linear) scale are larger for LX than L1 listeners in all cases except one (the difference between medium and hard noise in the audio-only condition). However, the proportional detriment of going from easier to harder noise (i.e., the change relative to identification in the easier noise level) was *smaller* for LX listeners in all cases, consistent with the interpretation of the logistic regression coefficients. This issue is quite common in logistic regression, and is driven by the fact that interaction effects are scale-dependent (see Rohrer & Arslan, 2021). Consistent with this claim, rebuilding the interaction model on the linear scale (i.e., assuming normally distributed residuals[6]) revealed a trend consistent with the condition means reported in Table 1: LX listeners were more negatively affected by noise than L1 listeners. Further, generating predictions based on the logistic regression model indicated larger effects of noise for L1 listeners when the outcome was predicted logits or odds, but larger effects of noise for LX listeners when the outcome was predicted probabilities. Thus, we refrain from making a broad claim about which group is "more affected" by noise, because the answer depends on whether one is interested in raw or proportional changes. However, it is worth nothing that the weaker modality effect in the LX sample persists on both the log-linear and linear scales.

Finally, the three-way interaction was significant, and examination of the summary output for the re-leveled model revealed that this effect was driven by the medium-hard comparison ($B_{hard*AV*LX} = -0.28$, $SE = 0.11$, $z = -2.54$, $p = .01$); neither individual three-way interaction coefficient estimate was significant in the model with easy as the reference level. Thus, we found evidence that the extent to which the talking face affects speech intelligibility across noise levels differed by participant group, and this effect is driven by the weaker modality-by-noise interaction for LX relative to L1 listeners *specifically for the medium-hard noise comparison*.

Subjective effort

The models reported below include 3,050 observations for L1 participants and 3,060 for LX participants. Ten observations were missing from L1 participants due to technical issues during data collection.

L1 Next we analyzed the effects of noise, modality, and their interaction on self-reported listening effort in the L1 group. In line with the intelligibility data, we found that participants reported less subjective effort in the audiovisual condition than the audio-only condition ($B_{AV} = -5.61$, $SE = 1.42$, $t = -3.97$, $p < .001$), controlling for noise level. In addition, subjective effort scores were lower in the easy condition than both the medium ($B_{medium} = 16.49$, $SE = 1.59$, $t = 10.39$, $p < .001$) and hard conditions ($B_{hard} = 37.71$, $SE = 2.30$, $t = 16.37$, $p < .001$), and lower in the medium relative to the hard condition ($B_{hard} = 21.22$, $SE = 1.69$, $t = 12.54$, $p < .001$), controlling for modality.

The interaction between noise and modality was significant, and the pattern of results mirrored that described above for the intelligibility data. The beneficial effect of seeing the talker was stronger in the hard than the medium SNR ($B_{AV*hard} = -6.02$, $SE = 1.17$, $t = -5.16$, $p < .001$), stronger in the medium than the easy SNR ($B_{AV*medium} = -5.75$, $SE = 1.17$, $t = -4.93$, $p < .001$), and stronger in the hard than the easy SNR ($B_{AV*hard} = -11.77$, $SE = 1.17$, $t = -10.08$, $p < .001$; Fig. 2).

[6] We are aware that using such models to analyze binary data violates regression assumptions and generates absurd predictions, but we ran this exploratory analysis for demonstration purposes.

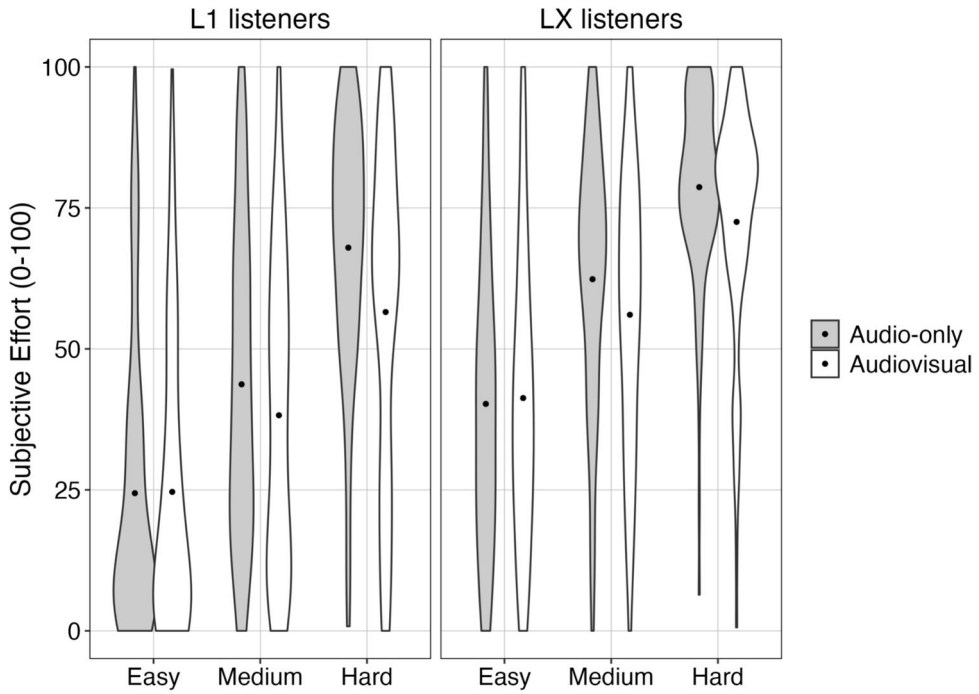

Fig. 2 Violin plot showing by-participant NASA-TLX ratings grouped by language group, noise, and modality. Dots indicate mean values for each condition

LX Subjective effort ratings were lower in the audiovisual condition than the audio-only condition ($B_{AV} = -3.81$, $SE = 1.04$, $t = -3.65$, $p < .001$), controlling for noise level. Ratings were also lower in the easy condition than both the medium ($B_{medium} = 18.45$, $SE = 1.54$, $t = 12.01$, $p < .001$) and hard conditions ($B_{hard} = 34.84$, $SE = 2.07$, $t = 16.86$, $p < .001$), and lower in the medium condition than the hard condition ($B_{hard} = 16.40$, $SE = 1.45$, $t = 11.29$, $p < .001$), controlling for modality. The interaction between noise and modality was also significant: Although the beneficial effect of seeing the talker on subjective effort ratings was not more pronounced in the hard relative to the medium SNR ($B_{AV*hard} = 0.14$, $SE = 1.10$, $t = 0.13$, $p = .90$), the modality effect was stronger in both the medium relative to the easy SNR ($B_{AV*medium} = -7.34$, $SE = 1.10$, $t = -6.67$, $p < .001$) and the hard relative to the easy SNR ($B_{AV*hard} = -7.20$, $SE = 1.10$, $t = -6.54$, $p < .001$).

Combined analysis Relative to L1 participants, LX participants reported more effort ($B = 13.25$, $SE = 2.60$, $t = 5.09$, $p < .001$). There were no group differences in the effects of modality or noise on subjective effort (see Table 3 above). The three-way interaction between participant group, noise, and modality was significant, and examination of the summary output for the model with the easy SNR as the reference level indicated that only the three-way interaction term involving the hard SNR was significant: The negative effect of noise on effort ratings at the hard SNR was less pronounced in the audiovisual condition than the audio-only condition, but this beneficial effect of seeing the talker in the hard noise level was less pronounced for LX participants ($B_{AV*hard*LX} = 4.57$, $SE = 1.61$, $t = 2.85$, $p = .004$). The three-way interaction term for the medium-hard comparison was also significant, again indicating that the beneficial effects of seeing the talker on effort ratings in the hard relative to the medium noise level was less pronounced for LX listeners ($B_{AV*hard*LX} = 6.16$, $SE = 1.60$, $t = 3.84$, $p < .001$).

Experiment 1 demonstrated that seeing the talker improves speech intelligibility and reduces subjective effort for both L1 and LX listeners, particularly in louder background noise. Relative to L1 listeners, LX listeners had lower accuracy and more subjective effort. Consistent with previous work (Drijvers & Özyürek, 2020; Xie et al., 2014; Yang et al., 2024), LX listeners showed less audiovisual intelligibility benefit than L1 listeners. Several other group differences emerged, including differences in the effect of noise on intelligibility and differences in the magnitude of the interaction between noise and modality on subjective effort.

Recall that the primary goal of this study was to evaluate the dual-task costs of audiovisual speech processing in L1 and LX listeners; Experiment 1 was included to obtain a measure of speech intelligibility in single-task conditions, and to replicate and build on some previous work addressing group differences in audiovisual speech processing and subjective effort. However, no work to date has addressed

whether L1 and LX listeners differ in the dual-task costs associated with processing audiovisual speech. Experiment 2 therefore implemented a dual-task paradigm to provide insight into how cognitive resources are allocated during native and nonnative speech processing.

Experiment 2

As described above, "listening effort" has been measured using a wide variety of paradigms (see, for example, Strand et al., 2018), and this variability persists even within the dual-task paradigm "umbrella" (see Gagné et al., 2017, for a review). However, many dual-task paradigms are not appropriate for audiovisual speech because the secondary task requires making visual judgments. To avoid auditory or visual interference, some studies have implemented tactile secondary tasks in which participants make speeded judgments to vibrotactile stimuli while listening to speech (e.g., Brown & Strand, 2019). Although these tasks can reliably detect changes in the dual-task costs associated with listening to speech in noise, they necessitate in-person data collection and require specialized equipment.

To overcome these issues, one of the authors recently developed and validated a novel dual-task paradigm to enable researchers to study the dual-task costs associated with audiovisual speech processing that can be implemented in online studies (Brown, 2025). This task is an auditory analogue to the vibrotactile task mentioned above, and requires that participants classify tones as short, medium, or long while listening to speech in noise. Crucially, the background noise is filtered to remove a band of frequencies approximately 200 Hz wide, centered on the frequency of the tone so that increasing the level of the background noise does not decrease the audibility of the tone. Therefore, any noise-induced changes in response times to the secondary tone task cannot be attributed to reduced audibility of the tone, so must instead reflect changes in the cognitive demands of the speech task (see Brown, 2025, for details).

Experiment 2 implemented this dual-task paradigm using new groups of L1 and LX participants. We used the same words and experimental conditions as in Experiment 1, but rather than simply identifying words, participants simultaneously completed the secondary tone classification task. Experiment 2 follows the conventions of Experiment 1 unless explicitly noted.

Method

Data, materials, and code for Experiment 2 are available online at https://osf.io/m29rw/. The preregistration document is available at https://osf.io/z7ert.

Participants

To reach our preregistered sample size of 120 participants per group, we recruited 136 L1 participants and 131 LX participants via Prolific. Before removing participants for meeting our preregistered exclusion criteria, we removed eight participants for not completing the task in at least one condition (i.e., identifying zero or one of the 50 words in a condition correctly). The decision to remove these individuals immediately—that is, before identifying participants to be removed due to any other exclusion criteria—was not preregistered (though excluding participants for having poor accuracy was preregistered). However, we opted to remove them because not completing the task means that their speech identification accuracies artificially deflate the average and increase variability (and therefore interfere with the data-driven exclusion criteria), which may also contaminate response times. We additionally removed individuals for meeting the following preregistered criteria: The participant's language background on Prolific did not match what they self-reported in the demographic questionnaire (one LX participant), accuracy at the speech identification task in any noise or modality condition was worse than three standard deviations below the mean (provided that the accuracy was also below 90%; L1: $N = 4$, LX: $N = 1$), mean response times were more than three standard deviations from the mean in any noise or modality condition (L1: $N = 4$, LX: $N = 3$), or mean tone classification accuracy was worse than three standard deviations below the mean in any noise or modality condition (one L1 participant). Together, these preregistered excursion criteria identified eight additional participants for exclusion. After excluding participants for these reasons, we had usable data from 125 L1 and 121 LX listeners; we only analyzed data from the first 120 usable participants in each group to adhere to our preregistered sample size.

Participants were compensated at a rate of $7 for 30 min of participation. Data collection occurred between December 20, 2024 and April 1, 2025.

Stimuli

Words and noise The word and noise stimuli were identical to those in Experiment 1.

Tones Each word trial contained a single 700-Hz tone, presented at the same level as the speech and lasting for 100 ms (short), 200 ms (medium), or 325 ms (long). The video files were edited to start and end with a closed mouth, resulting in an average word onset of 327 ms relative to the onset of the video file. Tones occurred either 400 ms or 750 ms after the onset of the video file. These positions were selected to prevent participants from predicting the onset of the tone

while ensuring that the tone overlapped with the word (or at least with the cognitive processing that occurs after the word that leads to word identification). Thus, on average, tones appeared either 73 ms or 423 ms after word onset (see stimulus spreadsheet at the OSF link for more information). Tone lengths and positions were yoked to words such that a given word was presented with the same tone at the same position in every condition. Each list of 50 words had approximately the same number of tone lengths and tone positions.

Procedure

After providing consent and passing the headphone screening, participants completed a tone exposure phase where they were familiarized with the tones they would hear in the main dual task to ensure they could identify the short, medium, and long tones. First, each tone length was played twice consecutively—the short tone twice, the medium tone twice, and then the long tone twice. Next, the sequence of three tones—short, medium, then long—was repeated three times. All tones in this phase were played without background noise.

After this phase, participants completed 18 randomized trials of a tone discrimination screening where they were asked to identify the tones they heard in the tone exposure phase using their keyboard, pressing "J" for short tones, "K" for medium tones, and "L" for long tones. After pressing the response key there was an interstimulus interval of 1,000 ms. If participants identified more than 12 tones correctly out of 18, they moved on to the next phase. If they identified fewer than 12 tones correctly, they were required to redo the tone discrimination screening. If they did not pass on the second screening, the experiment ended and they were given partial compensation. After successfully completing the discrimination screening, participants completed another 36 trials identical to the 18 trials in the previous phase. These trials were composed of twelve trials of each tone length presented at random, and were included because they provide us with a single-task measure of response times to the tone task (i.e., in the absence of speech) to use as a covariate in the analysis of the combined data from L1 and LX listeners (see the Combined Analyses section below). As in the tone discrimination screening, an interstimulus interval of 1,000 ms occurred between each trial. After these single-task tone trials, participants began the main task.

For the main task, participants heard each word in either the audio-only or audiovisual condition at one of three levels of background noise (easy, medium, or hard SNR). Each word contained either a short, medium, or long tone at the beginning or end of the word. The 300 words were divided into six lists. Given 50 words and three tone lengths, there were 16 words with medium tones and 17 each with short and long tones.

Participants were not instructed to prioritize one task over the other, but were required to complete the tasks in a specific order. They were told to first identify the tone lengths they heard in each word by pressing the corresponding keys used in the previous phases as quickly and accurately as possible. They were then prompted to type the word they heard in a text box. An interstimulus interval of 500 ms occurred between each of the 300 trials. After every list of 50 words, participants were given the opportunity to take a self-paced break.

Results and discussion

Unless otherwise noted, analyses followed the conventions outlined in Experiment 1. In all cases, participant random effects included random intercepts and random slopes for both noise and modality. None of the models including by-item (i.e., word) random slopes for noise converged. Thus, item random effects included random intercepts and by-word random slopes for modality for all models.

The flexible scoring method (Ponto; Kessler, 2017) corrected 5.4% of originally incorrect trials (see R script for implementation details). Mean accuracy at classifying the tones as short, medium, or long was 63.4%. The analyses reported below only include response time data from trials in which the tone was correctly classified (regardless of speech identification accuracy). Individual response times were excluded if the response time was more than three median absolute deviations (MADs) above or below that participant's median response time for that condition (Leys et al., 2013).[7] Given these exclusions, the models we report below include 21,910 observations for L1 participants and 20,928 for LX participants. The results of the model comparisons are shown in Tables 4 and 5, and coefficient estimates for each model are reported below in the text.[8]

L1

A likelihood ratio test indicated that the effect of background noise on response time was significant (see Table 4). Examination of the summary output for the model including effects of noise and modality indicated that response times did not differ significantly between

[7] Note that we preregistered that we would remove individual response times that were extreme for a particular participant in any noise or modality condition, but not any noise-by-modality condition. This was an error, however, as we intended to remove response time trials that were extreme for that participant in any condition.

[8] Although visual inspection of model assumptions revealed some evidence of non-normality and heteroskedasticity, we rebuilt key models with log-transformed response times as the outcome (which more closely adhered to model assumptions), and results were qualitatively the same.

Table 4 Results of model comparisons testing the effects of noise, modality, and their interaction on secondary task response times (ms) for the L1 and LX groups separately (Experiment 2)

Effect tested	Larger model (italicized term is omitted in the reduced model)	Group	Outcome of likelihood ratio test
Main effect of noise	*noise* + modality	L1	$\chi^2_2 = 32.65, p < .001$
		LX	$\chi^2_2 = 43.69, p < .001$
Main effect of modality	noise + *modality*	L1	$\chi^2_1 = 3.08, p = .08$
		LX	$\chi^2_1 = 1.68, p = .20$
Interaction between noise and modality	noise + modality + *noise:modality*	L1	$\chi^2_2 = 14.76, p < .001$
		LX	$\chi^2_2 = 40.66, p < .001$

Table 5 Results of model comparisons for secondary task response time data for the L1 and LX data combined (Experiment 2)

Effect tested	Larger model (italicized term is omitted in the reduced model)	Outcome of likelihood ratio test
Main effect of language group	noise + modality + single_task_rt + *group*	$\chi^2_1 = 7.67, p = .006$
Interaction between noise and language group	noise + modality + group + single_task_rt + *noise:group*	$\chi^2_2 = 5.09, p = .08$
Interaction between modality and language group	noise + modality + group + single_task_rt + *modality:group*	$\chi^2_2 = 0.18, p = .67$
Interaction among noise, modality, and language group	noise + modality + group + single_task_rt + noise:modality + noise:group + modality:group + *noise:modality:group*	$\chi^2_2 = 21.76, p < .001$

the easy and medium noise levels ($B_{medium} = 28.22$, $SE = 19.62$, $t = 1.44$, $p = .15$), but were an estimated 141 ms slower in the hard than the easy noise level ($B_{hard} = 140.79$, $SE = 23.90$, $t = 5.89$, $p < .001$). Re-leveling the noise variable with "medium" rather than "easy" as the reference level revealed that response times in the hard noise level were on average an estimated 113 ms slower in the hard than the medium noise level ($B_{hard} = 112.58$, $SE = 22.01$, $t = 5.12$, $p < .001$). The effect of modality was not significant ($B_{modality} = -31.14$, $SE = 17.69$, $t = -1.76$, $p = .08$), but the interaction between noise and modality significantly improved model fit. The reduction in dual-task costs associated with seeing the talker was more pronounced in the hard than the easy SNR ($B_{hard*AV} = -62.92$, $SE = 17.92$, $t = -3.51$, $p < .001$; in fact, the estimate for the modality effect was numerically positive in the easy SNR) and more pronounced in the medium than the easy SNR ($B_{medium*AV} = -53.53$, $SE = 17.55$, $t = -3.05$, $p = .002$), but did not differ between the medium and the hard SNR ($B_{hard} = 9.39$, $SE = 18.12$, $t = -0.52$, $p = .60$). Thus, consistent with previous work (Brown, 2025), seeing the talker only reduced dual-task costs when the addition of the visual signal was necessary to understand the speech—that is, in moderate and difficult levels of background noise, but not easy ones (see Fig. 3).

LX

The data from the LX listeners mirror those described above for the L1 listeners. Controlling for modality, secondary task response times differed by noise level, with response times on average an estimated 79 ms slower in medium ($B_{medium} = 79.10$, $SE = 20.46$, $t = 3.87$, $p < .001$) and 221 ms slower in hard levels of background noise ($B_{hard} = 220.74$, $SE = 31.21$, $t = 7.07$, $p < .001$) relative to the easy condition. Re-leveling the noise variable revealed that response times were an estimated 142 ms slower in hard relative to medium noise ($B_{hard} = 141.64$, $SE = 30.72$, $t = 4.61$, $p < .001$). As in the L1 analysis, the effect of modality was not significant for LX listeners (controlling for noise level), but the interaction between noise and modality significantly improved model fit. In general, there was a greater reduction in dual-task costs in more difficult listening conditions, consistent with the results from the L1 listeners. However, for the LX participants, the reduction in dual-task costs associated with seeing the talker was significantly more pronounced in the hard than the medium SNR ($B_{hard*AV} = -129.94$, $SE = 20.78$, $t = -6.25$, $p < .001$) and more pronounced in the hard than the easy SNR ($B_{hard*AV} = -90.77$, $SE = 20.59$, $t = -4.41$, $p < .001$), but did not differ between the medium and easy SNRs ($B_{medium*AV} = 39.16$, $SE = 20.00$, $t = 1.96$, $p = .05$).

Note that the numerically positive interaction term in the medium noise level is somewhat puzzling and inconsistent with the theoretical framework that seeing the talker relieves cognitive load to a greater extent as listening conditions become more difficult. This effect is nonsignificant, but is likely driving the significant three-way interaction between language group, noise, and modality described below (see Fig. 3; Table 5).

Combined analyses

The final set of analyses combined the data from L1 and LX listeners. We first assessed whether L1 and LX participants differed in how quickly they responded to the single-task trials. If we were to find group differences in response times to the tone classification task when no speech was presented, it would confound the claim that group differences in the dual task trials reflect the cognitive costs of listening to speech. The single-task analysis revealed no significant effect of language background when speech was *not* present ($\chi^2_1 = 1.49$, $p = .22$). In fact, response times were numerically slower for L1 listeners than LX listeners ($B = -24.80$, $SE = 20.39$, $t = -1.22$, $p = .23$). Thus, it is clearly not the case that individuals who learned another language before learning English have slower response times overall; instead, these results suggest that listening to speech in one's nonnative language is more cognitively demanding than listening to speech in one's native language.

For the analysis of the dual-task trials that included language background below, we calculated the mean single-task response time for each participant—excluding incorrect trials to be consistent with the dual-task analyses—and included this variable in the model to control for individual differences in processing speed and/or hardware differences that might affect overall response times. A likelihood ratio test indicated that the three-way interaction between noise, modality, and participant group was significant (see Table 4). Examination of the summary output for the model including the three-way interaction indicated that the three-way interaction was significant for the medium but not the hard noise level. This effect was driven by the numerically positive (though nonsignificant) interaction term for the LX listeners in the medium noise level. Indeed, this effect is inconsistent with the theoretical framework described in the Introduction section—informed by previous research (Brown, 2025; Brown & Strand, 2019)—and with the general trend that the beneficial effects of audiovisual speech on dual-task costs become more pronounced as listening conditions become more difficult. To statistically evaluate the claim that the three-way interaction is driven by the medium noise level, we conducted exploratory analysis in which we only analyzed data from the easy and hard conditions, which revealed a nonsignificant three-way interaction ($\chi^2_1 = 1.05$, $p = .31$). Thus overall, we have some evidence that the magnitude of the noise-by-modality interaction may differ by language background, but given the puzzling nature

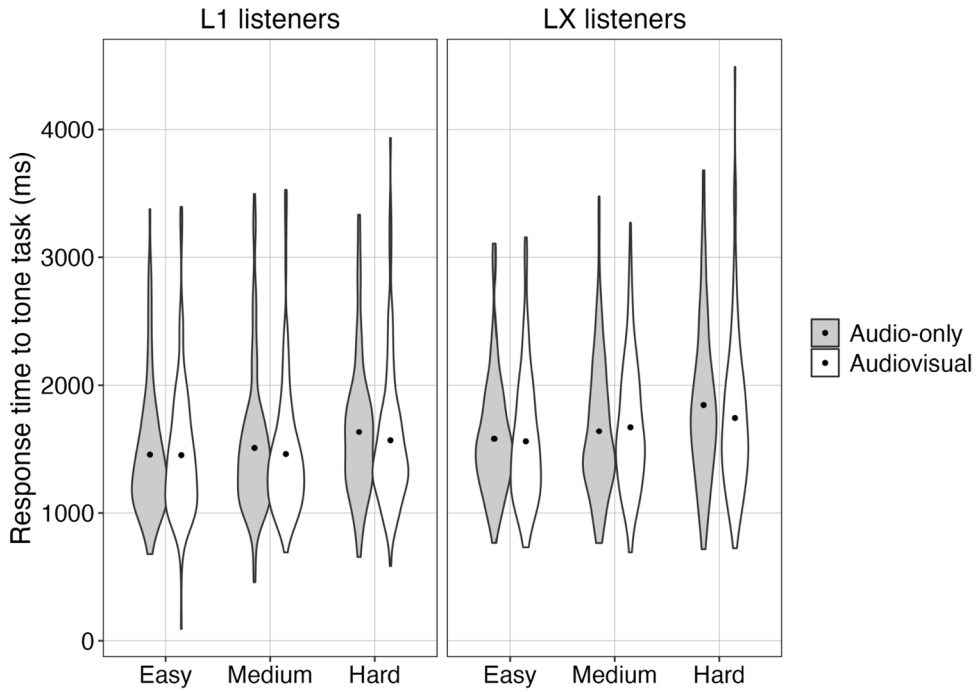

Fig. 3 Violin plot showing by-participant secondary task response times grouped by language group, noise, and modality. Dots indicate mean values for each condition

of the interaction, this finding should be replicated in future research. Neither of the two-way interactions between participant group and noise or modality was significant. However, the main effect of language background was significant such that response times for LX listeners were on average an estimated 153 ms slower than L1 listeners ($B = 152.95$, $SE = 54.50$, $t = 2.81$, $p = .005$).

General discussion

Experiment 1 revealed clear effects of background noise as well as audiovisual intelligibility benefits for both L1 and LX listeners. As expected, noise impaired speech intelligibility, and identification accuracy improved when participants could both see and hear the talker relative to audio-only conditions. Consistent with prior work on word identification, this audiovisual benefit became more pronounced as the level of the background noise increased (e.g., Brown & Strand, 2019; Ross et al., 2006; Sumby & Pollack, 1954); indeed, the visual signal has greater opportunity to improve speech identification when audio-only accuracy is low.[9] Thus, LX listeners can benefit from the phonetic cues provided by a talking face, despite having less experience with the language.

Although all effects of interest (i.e., noise, modality, and the noise-by-modality interaction) emerged in both groups, we found some differences in the magnitudes of these effects across groups. For intelligibility, we found that LX listeners showed less audiovisual benefit and were differently affected by background noise relative to L1 listeners (though assigning a direction to the effect is more complicated than it may appear; see above). We also found some evidence that the magnitude of the noise-by-modality interaction differed by group, though this effect only emerged for one of the three pairwise comparisons. For subjective effort, the only group difference was in the magnitude of the noise-by-modality interaction. Finally, LX listeners had lower speech identification accuracy and rated the speech identification task as more effortful overall relative to L1 listeners.

Experiment 2 used a dual-task paradigm to assess the cognitive costs associated with processing noisy speech in audio-only and audiovisual conditions. As in Experiment 1, audiovisual speech led to better word identification accuracy than audio-only speech for both L1 and LX participants (see Table 3), even under dual-task demands. This is to be expected and suggests that visual speech cues continue to support speech identification when attention is divided. Surprisingly, however, examination of Table 1 indicates that speech identification accuracy was considerably *better* in Experiment 2 than in Experiment 1, particularly as the listening conditions became more difficult. This is unexpected given the assumption that dividing attention between two challenging tasks uses resources that exist in finite amounts (Kahneman, 1973), thereby impairing performance on both tasks. Further, given the assumption of multiple resource theory that tasks that are similar in terms of modality or other overlapping features (see Brown, 2025; Isreal et al., 1980) interfere with one another to a greater extent than unrelated tasks in separate modalities, one might expect substantial interference from the auditory secondary task that would impair performance on the primary task. This puzzling finding may perhaps suggest that a challenging secondary task—particularly one that does not interfere with sensory processing in either of the primary task modalities—leads to greater task engagement and therefore better performance on the primary task. This may allay the concern about dual-task paradigms that they lead to poorer primary task performance, which complicates interpretation of dual-task costs. However, we hesitate to draw firm conclusions from this trend, and suggest that future researchers attempt to clarify the situations in which dual-tasking may actually improve primary task performance.

The outcome of greatest interest in Experiment 2, however, was response time to the secondary tone classification task. Consistent with a large body of work in the listening effort literature, increasing the level of the background noise increased cognitive load in both groups, as indicated by slower response times to the secondary tone task. This finding is consistent with the claim that more challenging listening conditions demand greater cognitive resources, leaving fewer resources available to quickly complete the challenging tone classification task (e.g., Brown & Strand, 2019; Gagné et al., 2017; Picou & Ricketts, 2014). The interaction between modality and noise level also emerged in both groups, and the effect was in the hypothesized direction based on prior research (Brown, 2025; Brown & Strand, 2019): Seeing the talker sped secondary task response times (relative to the audio-only condition) *only in difficult listening conditions*. These results suggest that although processing speech in two modalities might incur a small processing cost—perhaps as a result of integration costs, simultaneously monitoring two channels, distraction, or some other mechanism—the reduction in lexical competition afforded by visual phonetic cues offsets these costs when the listening conditions are challenging.

Although we did not find clear evidence of an audiovisual processing cost in this study (i.e., response times did not differ between audio-only and audiovisual

[9] Note, however, that this pattern may differ for sentence-length materials; indeed, `Xie et al. (2014) showed that audiovisual benefit decreased at SNRs below −12. This may occur because sentence-length materials are very difficult to lipread, especially for listeners who are not highly familiar with the language.

conditions in the easy SNR), the direction and magnitude of the effect of seeing the talker on dual-task costs is dependent upon the relative difficulty of the speech task in the "easy" listening condition, as well as the difficulty of the secondary task (see Brown, 2025). Together, these results reinforce the notion that although audiovisual speech is beneficial in adverse listening conditions—whether "beneficial" refers to intelligibility or dual-task costs—visual cues may not always reduce cognitive load relative to audio-only conditions, particularly when such cues are unnecessary to successfully identify the speech. Overall, the fact that the effects of noise, modality, and their interaction on dual-task costs were robust and in the same direction for both language groups suggests that the cognitive mechanisms underlying audiovisual speech processing function similarly for individuals with less experience with the language.

Experiment 2 also revealed that secondary task response times were slower overall for LX listeners, suggesting that processing speech in one's nonnative language is more cognitively demanding than processing speech in one's native language (see Borghini & Hazan, 2018; Peng & Wang, 2019). Crucially, this effect cannot be driven by baseline differences in response times between groups: Analysis of single-task response time data indicated that response times to classify the tone in noise did not differ between L1 and LX listeners. Although LX listeners had increased dual-task costs overall, there was no evidence that the magnitude or direction of the noise or modality effects on dual-task costs differed by language group. These findings—namely, that LX listeners expend greater cognitive effort to identify speech in their nonnative language than L1 listeners, but this effect does not appear to be moderated by the difficulty of the listening task—are consistent with prior work using pupillometry to measure listening effort (Borghini & Hazan, 2018). Together, results of the experiments reported here suggest that although background noise level and modality may have different effects on speech intelligibility for L1 and LX listeners (Experiment 1), these variables do not appear to differentially affect cognitive demand (Experiment 2).

The only interaction with language background that emerged in Experiment 2 was the three-way interaction between noise, modality, and language background, which provides some evidence that the magnitude of the interaction between noise and modality may differ for L1 and LX listeners. However, as we mentioned in the "Results and Discussion" section of Experiment 2, this interaction appears to be driven by an unexpected increase in mean response times for audiovisual speech in the medium noise level, which only occurred for LX listeners. Indeed, when we omitted the data from the medium SNR and only analyzed the data from the easy and hard conditions, we did not find evidence that the magnitude of the noise-by-modality interaction differed by language background. Thus, audiovisual speech appears to similarly affect dual-task costs for L1 and LX listeners at most noise levels, though there may be some differences at moderate levels of listening difficulty. This could be because a given SNR might not lead to the same actual level of difficulty across groups; therefore, the more informative data points are the general changes across a range of difficulty rather than pairwise comparisons. Future researchers might consider conducting a similar study to this one that includes a wider range of SNRs and perhaps smaller changes between adjacent SNRs to obtain a more fine-grained picture of the relationship between modality, noise, and dual-task costs for L1 and LX listeners.

Constraints on generality

The current study demonstrated similar patterns of results regarding the effects of modality and background noise level on speech intelligibility, subjective listening effort, and dual-task costs. As we discuss in the paper, modality effects are highly dependent on background noise level (regardless of the outcome measure). We therefore would not necessarily expect the effect of seeing the talker to be the same as we report here at different levels of background noise, or different manipulations of task difficulty (e.g., reverberation, sine wave speech). However, we expect the general pattern of increased audiovisual benefit on response times in more difficult listening conditions to persist for most manipulations of task difficulty. This experiment also used isolated words, so we do not wish to make claims about the extent to which seeing the talker affects dual-task costs for sentence- or passage-length materials, though this is a fruitful avenue for future research. Our LX listeners included individuals with a wide variety of L1s, so we do not expect that these results are dependent on the particular languages spoken by LX participants. However, participants must have either self-reported or objectively-demonstrated fluency in English for these results to apply. Finally, our participants had self-reported normal hearing and normal or corrected-to-normal vision and were between 18 and 45 years of age; we do not necessarily expect these results to generalize to much older or much younger samples, nor to individuals with hearing or uncorrected visual limitations.

Conclusions

The most notable takeaway from this study was that although there were some group differences in the *magnitudes* of the effects (more so for intelligibility than dual-task costs), all effects of interest across both experiments were robust and in the same direction for L1 and LX listeners. Indeed, effects

of modality, background noise level, and their interaction persisted in both groups, regardless of whether the outcome was speech intelligibility, subjective ratings of listening effort, or dual-task costs. Together, these data suggest that if researchers are interested in establishing general findings regarding audiovisual speech processing across levels of background noise, including LX listeners like those sampled here are not likely to significantly affect outcomes.[10] These results are consistent with our recent work demonstrating that effects of lexical difficulty and semantic constraint on intelligibility and subjective listening effort persist regardless of language background (Strand et al., 2024). The current study therefore adds to the growing list of general speech-related phenomena that emerge in listeners coming from a wide variety of language backgrounds, and provides further evidence that loosening or eliminating the "native English speaker" participation restriction often does not substantially impact study outcomes (and has the potential to diversify our samples and therefore increase the generalizability of our work; see Ghai et al., 2025 for more on the value of including more diverse samples in psychological research). To be clear: We are not arguing that the entire field of psycholinguistics and related domains should ignore language background, nor that studying differences in speech processing between L1 and LX listeners is not valuable work. Quite the contrary. Rather, our goal is to encourage researchers to reflect on whether and how participant selection might affect study outcomes, and whether limiting samples to L1 listeners is likely to be worth the reduction in generalizability that comes with sample homogeneity.

Authors' note Portions of this project were presented at the annual meeting of the Midwest Psychology Association and at the Minnesota Undergraduate Psychology Conference (2025).

Funding This work was supported by Carleton College and National Institute on Deafness and Other Communication Disorders via a grant to Julia Strand (R15-DC018114).

Data availability All data, materials, and analysis scripts are available online (https://osf.io/m29rw/).

Code availability All code are publicly available online (https://osf.io/m29rw/).

Declarations

Conflicts of interest/Competing interests None.

Ethics approval The Carleton College Institutional Review Board approved all research procedures.

Consent to participate Participants consented to participate.

Consent for publication We consent.

[10] It is also worth noting that the groups included in the current study consisted of 100% L1 or 100% LX listeners, whereas loosening the "native English speaker" participation criterion would be unlikely to produce a homogenous LX sample in practice. Thus, any group differences reported here would likely have little impact on study conclusions in a heterogeneous sample.

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

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, 52(1), 388–407.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R., Singmann, H., & Green, P. (2014). *Package "lme4"* (Versions 1.1-15) [Computer software]. R foundation for Statistical Computing. https://github.com/lme4/lme4/

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*(2), 276–292.

Black, J. W., & Hast, M. H. (1962). Speech reception with altering signal. *Journal of Speech and Hearing Research, 5*, 70–75.

Blackburn, C. L., Kitterick, P. T., Jones, G., Sumner, C. J., & Stacey, P. C. (2019). Visual speech benefit in clear and degraded speech depends on the auditory intelligibility of the talker and the number of background talkers. *Trends in Hearing, 23*, 2331216519837866.

Borghini, G., & Hazan, V. (2018). Listening effort during sentence processing is increased for non-native listeners: A pupillometry study. *Frontiers in Neuroscience, 12*, 152.

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America, 106*(4 Pt 1), 2074–2085.

Brown, V. A. (2025). Measuring the dual-task costs of audiovisual speech processing across levels of background noise. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001826

Brown, V. A., & Strand, J. F. (2019). About face: Seeing the talker improves spoken word recognition but increases listening effort. *Journal of Cognition*. https://doi.org/10.5334/joc.89

Brown, V. A., McLaughlin, D. J., Strand, J. F., & Van Engen, K. J. (2020). Rapid adaptation to fully intelligible nonnative-accented speech reduces listening effort. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/1747021820916726

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.

Cheng, L. S. P., Burgess, D., Vernooij, N., Solís-Barroso, C., McDermott, A., & Namboodiripad, S. (2021). The problematic concept of native speaker in psycholinguistics: Replacing vague and harmful terminology with inclusive and accurate measures. *Frontiers in Psychology, 12*, Article 715843.

Cooke, M., Garcia Lecumberri, M. L., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *The Journal of the Acoustical Society of America, 123*(1), 414–427.

Dewaele, J.-M. (2018). Why the dichotomy "L1 versus LX user" is better than "native versus non-native speaker." *Applied Linguistics, 39*(2), 236–240.

Drijvers, L., & Özyürek, A. (2020). Non-native listeners benefit less from gestures and visible speech than native listeners during

degraded speech comprehension. *Language and Speech, 63*(2), 209–220.
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.
Francis, A. L., & Love, J. (2020). Listening effort: Are we measuring cognition or affect, or both? *Wiley Interdisciplinary Reviews. Cognitive Science, 11*(1), Article e1514.
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*, 1–25.
Ghai, S., Thériault, R., Forscher, P., Shoda, Y., Syed, M., Puthillam, A., Peng, Hu Chuan, Basnight-Brown, Dana, Majid, Asifa, Azevedo, Flavio, & Singh, L. (2025). A manifesto for a globally diverse, equitable, and inclusive open science. *Communications Psychology, 3*(1), Article 16.
Grant, K. W., & Walden, B. E. (1996). Evaluating the articulation index for auditory-visual consonant recognition. *The Journal of the Acoustical Society of America, 100*(4), 2415–2424.
Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America, 103*(5 Pt 1), 2677–2690.
Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology, 52*, 139–183.
Herrmann, B., & Johnsrude, I. S. (2020). A model of listening engagement (MoLE). *Hearing Research, 397*, Article 108016.
Hintz, F., Voeten, C. C., & Scharenborg, O. (2023). Recognizing non-native spoken words in background noise increases interference from the native language. *Psychonomic Bulletin & Review, 30*(4), 1549–1563.
Isreal, J. B., Chesney, G. L., Wickens, C. D., & Donchin, E. (1980). P300 and tracking difficulty: Evidence for multiple resources in dual-task performance. *Psychophysiology, 17*(3), 259–273.
Kahneman, D. (1973). *Attention and effort*. Prentice-Hall.
Kessler, B. (2017). *Ponto*. http://spell.psychology.wustl.edu/ponto/
Kramer, S. E., Lorens, A., Coninx, F., Zekveld, A. A., Piotrowska, A., & Skarzynski, H. (2012). Processing load during listening: The influence of task characteristics on the pupil response. *Language and Cognitive Processes, 28*(4), 426–442.
Kuchinsky, S. E., Ahlstrom, J. B., Vaden, K. I., Jr., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology, 50*(1), 23–34.
Kuchinsky, S. E., Gallun, F. J., & Lee, A. K. C. (2024). Note on the dual-task paradigm and its use to measure listening effort. *Trends in Hearing, 28*, 23312165241292216.
Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764–766.
Mackersie, C. L., & Cones, H. (2011). Subjective and psychophysiological indexes of listening effort in a competing-talker task. *Journal of the American Academy of Audiology, 22*(2), 113–122.
Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research: JSLHR, 50*(4), 940–967.
McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group "white paper." *International Journal of Audiology, 53*(7), 433–445.
Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22–27.
Peng, Z. E., & Wang, L. M. (2019). Listening effort by native and nonnative listeners due to noise, reverberation, and talker foreign accent during English speech perception. *Journal of Speech, Language, and Hearing Research: JSLHR, 62*(4), 1068–1081.
Picou, E. M., Gordon, J., & Ricketts, T. A. (2016). The effects of noise and reverberation on listening effort in adults with normal hearing. *Ear and Hearing, 37*(1), 1–13.
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.
R Core Team. (2022). *R (Version 4.2.2) [Computer software]*. R Foundation for Statistical Computing.
Rabbitt, P. M. (1968). Channel-capacity, intelligibility and immediate memory. *Quarterly Journal of Experimental Psychology, 20*(3), 241–248.
Rohrer, J. M., & Arslan, R. C. (2021). Precise answers to vague questions: Issues with interactions. *Advances in Methods and Practices in Psychological Science, 4*(2), 25152459211007370.
Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2006). Do you see what i am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex, 17*(5), 1147–1153.
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.
Scharenborg, O., & van Os, M. (2019). Why listening in background noise is harder in a non-native language than in a native language: A review. *Speech Communication, 108*, 53–64.
Seeman, S., & Sims, R. (2015). Comparison of psychophysiological and dual-task measures of listening effort. *Journal of Speech, Language, and Hearing Research: JSLHR, 58*(6), 1781–1792.
Sommers, M. S., & Phelps, D. (2016). Listening effort in younger and older adults: A comparison of auditory-only and auditory-visual presentations. *Ear and Hearing, 37*(Suppl 1), 62S-68S.
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, 61*, 1463–1486.
Strand, J. F., Ray, L., Dillman-Hasso, N. H., Villanueva, J., & Brown, V. A. (2021). Understanding speech amid the jingle and jangle: Recommendations for improving measurement practices in listening effort research. *Auditory Perception & Cognition, 3*(4), 1–20.
Strand, J. F., Brown, V. A., Sewell, K., Lin, Y., Lefkowitz, E., & Saksena, C. G. (2024). Assessing the effects of "native speaker" status on classic findings in speech research. *Journal of Experimental Psychology. General*. https://doi.org/10.1037/xge0001640
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.
Vulchanova, M., Vulchanov, V., Sorace, A., Suarez-Gomez, C., & Guijarro-Fuentes, P. (2022). Editorial: The notion of the native speaker put to the test: Recent research advances. *Frontiers in Psychology, 13*, Article 875740.
Wagner, A. E., Toffanin, P., & Başkent, D. (2016). The timing and effort of lexical access in natural and degraded speech. *Frontiers in Psychology, 7*, Article 398.

Winn, M. B. (2018). *Praat script for creating speech-shaped noise* (Version 12) [Computer software]. http://www.mattwinn.com/praat.html

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics, 79*(7), 2064–2072.

Xie, Z., Yi, H.-G., & Chandrasekaran, B. (2014). Nonnative audiovisual speech perception in noise: Dissociable effects of the speaker and listener. *PLoS One, 9*(12), Article e114439.

Yang, J., Nagaraj, N. K., & Magimairaj, B. M. (2024). Audiovisual perception of interrupted speech by nonnative listeners. *Attention, Perception, & Psychophysics*. https://doi.org/10.3758/s13414-024-02909-3