Attention, Perception, & Psychophysics
https://doi.org/10.3758/s13414-022-02486-3

Revisiting the target-masker linguistic similarity hypothesis

Violet A. Brown[1] · Naseem H. Dillman-Hasso[2] · ZhaoBin Li[2] · Lucia Ray[2] · Ellen Mamantov[2] · Kristin J. Van Engen[1] · Julia F. Strand[2]

Accepted: 25 March 2022

Abstract

The linguistic similarity hypothesis states that it is more difficult to segregate target and masker speech when they are linguistically similar. For example, recognition of English target speech should be more impaired by the presence of Dutch masking speech than Mandarin masking speech because Dutch and English are more linguistically similar than Mandarin and English. Across four experiments, English target speech was consistently recognized more poorly when presented in English masking speech than in silence, speech-shaped noise, or an unintelligible masker (i.e., Dutch or Mandarin). However, we found no evidence for graded masking effects—Dutch did not impair performance more than Mandarin in any experiment, despite 650 participants being tested. This general pattern was consistent when using both a cross-modal paradigm (in which target speech was lipread and maskers were presented aurally; Experiments 1a and 1b) and an auditory-only paradigm (in which both the targets and maskers were presented aurally; Experiments 2a and 2b). These findings suggest that the linguistic similarity hypothesis should be refined to reflect the existing evidence: There is greater release from masking when the masker language differs from the target speech than when it is the same as the target speech. However, evidence that unintelligible maskers impair speech identification to a greater extent when they are more linguistically similar to the target language remains elusive.

Keywords Linguistic similarity hypothesis · Speech identification · Masking · Cross-modal masking

Introduction

Processing spoken language is often complicated by the presence of simultaneous conversations, environmental sounds, and other types of noise. Background noise can impair a listener's ability to recognize target speech through both low-level sensory and higher-level cognitive mechanisms (e.g., see Freyman et al., 1999; Kidd et al., 1998). The former, known as *energetic masking*, interferes at the level of the auditory periphery such that the intelligibility of the target speech is impaired by background noise as a result of spectral and temporal overlap between the acoustic signals. The other broad class of masking is referred to as *informational masking* (previously known as "perceptual masking"; Carhart et al., 1969), which describes any masking that is not energetic (e.g., difficulty attending to the target stream, stimulus uncertainty, etc.; see Agus et al., 2009). Distinguishing between the effects of energetic and informational masking can help to clarify the perceptual and cognitive processes underlying spoken word recognition.

One mechanism that may underlie the challenges associated with informational masking is that perceptual similarity between the target and masker may lead to difficulty with stream segregation (Brungart et al., 2001; Calandruccio et al., 2013; Durlach et al., 2003; Festen & Plomp, 1990). Masking is particularly detrimental to speech recognition when the target and speech masker are confusable; for example, listeners have more difficulty identifying words when the masker comes from the same location as the target (Freyman et al., 2001; Helfer et al., 2010; Ihlefeld & Shinn-Cunningham, 2008; Rothpletz et al., 2012) and when the masker and target are produced by same-sex talkers (Brungart et al., 2001; Festen & Plomp, 1990). Confusability that leads to errors of stream segregation need not be limited to low-level grouping cues like similarity in frequency (Bregman & Campbell, 1971) or temporal synchrony (Dannenbring & Bregman, 1978). Given that listeners can successfully segregate speech streams despite considerable overlap in frequency

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

[1] Department of Psychological and Brain Sciences, Washington University in St. Louis, One Brookings Drive, St. Louis, MO 63130, USA

[2] Carleton College, Department of Psychology, One North College St, Northfield, MN 55057, USA

Published online: 26 April 2022

ranges across talkers (Newman & Evers, 2007), higher-level non-acoustic cues—such as familiarity with the talker (Johnsrude et al., 2013) or even the language itself (Van Engen & Bradlow, 2007)—also play a role in stream segregation.

The *target-masker linguistic similarity hypothesis* (sometimes referred to simply as the linguistic similarity hypothesis) predicts that listeners will have greater difficulty distinguishing between the target and masker—and therefore impaired identification of the target speech—when the target and masker are similar to one another (Brouwer et al., 2012). In support of the linguistic similarity hypothesis, English-speaking listeners identify English speech more poorly when the maskers are English relative to Spanish (Garcia Lecumberri & Cooke, 2006), Dutch (Brouwer et al., 2012; Freyman et al., 2001), or Mandarin (Van Engen & Bradlow, 2007). The detrimental effects of target-masker similarity have also been shown for regional variations of languages (Brouwer, 2017) and accented speech (Brouwer, 2019; Calandruccio et al., 2010). These results are consistent with the claim that stream segregation is more challenging when targets and maskers are linguistically similar to each other, but it is not currently clear what features of target-masker similarity drive these effects. That is, similar languages may provide more interference because of similarity in acoustic features, temporal patterns, or prosody. Alternatively, the interference may be higher-level—maskers that are more similar to the target may attract more attention and therefore lead to poorer identification of the target speech. Interference may occur on the level of the whole word, as some studies have reported that participants transcribe entire words from the masker (Summers & Roberts, 2020), but it may also reflect sublexical or prosodic differences between the maskers (see Calandruccio et al., 2019; Van Engen & Bradlow, 2007). Regardless of the precise mechanism, the linguistic similarity hypothesis rests on the assumption that maskers that are more similar to the target provide greater *informational* masking than those that are less similar.[1]

Isolating the unique influence of informational masking—in research on the linguistic similarity hypothesis and elsewhere—is not a trivial task (but see Kidd et al., 2005; Summers & Roberts, 2020). One reason is that maskers that are thought of as primarily informational, such as a single-talker masker or two-talker babble, typically also overlap spectrally with the target speech, resulting in energetic masking as well (Brungart et al., 2001). This makes it difficult to dissociate higher-level cognitive interference (informational masking) from lower-level acoustic similarity (energetic masking). Indeed, although informational and energetic masking are typically discussed separately, they often occur simultaneously and are therefore difficult to disentangle. For example, though listeners can more effectively segregate the target and competing speech streams when they are produced by different-sex talkers relative to same-sex talkers (Brungart et al., 2001; Festen & Plomp, 1990; Williams & Viswanathan, 2020), these differences are partially attributable to energetic masking because same-sex voices have greater spectral overlap than different-sex voices.

The challenge of disentangling energetic and informational masking is apparent in the literature on the linguistic similarity hypothesis. For example, Calandruccio et al. (2013) presented native English speakers with English target speech in the presence of English, Dutch (which is linguistically similar to English), or Mandarin (which is less similar) two-talker babble (see Bradlow et al., 2010, for a discussion and classification of phonetic similarity across languages). Calandruccio and colleagues predicted that recognition of the target speech would be most impaired by English masking, less impaired by Dutch, and least impaired by Mandarin. The analyses showed that the difference in performance between the English and Mandarin masking conditions was significantly larger than the difference between English and Dutch masking, in line with the predictions of the linguistic similarity hypothesis (see also Brouwer et al., 2012; Van Engen, 2010; Van Engen & Bradlow, 2007). However, a follow-up experiment showed that the three language maskers used in that study had different spectral properties, and these spectral differences were partially responsible for the observed effects (Calandruccio et al., 2013).

One way that studies have tried to reduce differences in energetic masking across languages is by matching the long-term average spectrum (LTAS) of the different language maskers. For example, Brouwer and colleagues (Brouwer et al., 2012) matched English and Dutch maskers on LTAS and showed that listeners still experienced more masking when the target and masker were presented in the same language than when they were presented in different languages, even when the LTAS of the two maskers did not differ. Although matching on LTAS helps to reduce the possibility that differences across languages are spectral in nature, LTAS matching does not match spectral properties on a moment-to-moment basis, so it does not completely remove differences in energetic masking between languages. That is, after LTAS matching, there may still be acoustic differences between the maskers such as greater short-term changes in pitch, differences in syllable rate, or differences in mean fundamental frequency. Even if LTAS matching is performed across sentences, it is possible that the extent of energetic masking at a given moment in time is influenced by the temporal distribution of frequencies within the target and masker, which is unaffected by LTAS matching and may differ across languages.

[1] Note that the predictions of the linguistic similarity hypothesis here are also in line with research on the irrelevant sound effect (Jones & Macken, 1993). For example, maskers in a participant's native language impair visual working memory more than maskers in an unfamiliar language (Ellermeier et al., 2015).

The strongest test of the linguistic similarity hypothesis—and the more general claim that informational masking impairs intelligibility above and beyond the effects of energetic masking—would require that the influence of energetic masking be completely removed so there is no spectral overlap between the target and masker. One way to isolate the effects of informational masking would be to ensure that the target speech and background noise occupy different frequency bands (e.g., see Agus et al., 2009; Kidd et al., 2005), so that any differences across the background languages could not be attributable to varying degrees of spectral overlap with the target. However, this technique requires substantial manipulation of the speech spectra, and different languages have different spectral characteristics (Byrne et al., 1994), so this is not a feasible technique for studying differences across languages. Another method of isolating informational masking would be to present the auditory target and masker in different ears. For example, Summers and Roberts (2020) demonstrated that intelligible maskers provide greater interference than acoustically-similar unintelligible maskers, even when energetic masking effects were removed by presenting targets and maskers to opposite ears. However, dichotic presentation provides challenges for testing linguistic similarity effects, as those effects are attenuated when the target and masker are presented in different spatial locations (Viswanathan et al., 2016).

Another way to isolate informational masking and completely remove the influence of energetic masking is to present the target speech and the background babble in different modalities: Participants hear masking noise while lipreading the target speech (Campbell et al., 2002;Lidestam et al., 2014 ; Myerson et al., 2016). Prior work on cross-modal masking has suggested that high-level cognitive mechanisms contribute to masking even in the absence of sensory interference. When participants perform a lipreading task in the presence of multi-talker babble, they perform more poorly than when lipreading in silence or in steady-state noise (Lidestam et al., 2014; Myerson et al., 2016). Importantly, in both previous studies, lipreading performance did not differ in silence and steady-state noise, suggesting that it was not simply the presence of any noise that interfered with lipreading, but rather the presence of speech specifically.[2]

The finding that two-talker babble impairs lipreading suggests that the detrimental effects of informational masking are not limited to the auditory modality, and that cross-modal paradigms can be used to isolate the effects of informational masking in speech perception research. The use of a cross-modal paradigm to test the linguistic similarity hypothesis is further supported by strong parallels between auditory and visual speech perception in other domains (see Rosenblum, 2008). For example, word recognition in both modalities is similarly affected by lexical characteristics such as neighborhood density and word frequency (Strand & Sommers, 2011). We opted to use a cross-modal masking paradigm because it removes the influence of energetic masking effects. Although this enables us to distinguish between competing explanations for why the linguistic similarity effect occurs, using visual rather than auditory targets changes some of the task demands. For example, the difficulties of stream segregation may be alleviated when the target and masker are in different modalities. However, demonstrating effects consistent with the linguistic similarity hypothesis when the target and maskers are presented in different modalities would provide strong support for the claim that linguistic similarity between languages—in addition to low-level spectral overlap—is responsible for the observed interference.

In Experiment 1a, we assessed native English speakers' ability to lipread English sentences in six conditions: silence, speech-shaped noise, English (meaningful), English (anomalous), Dutch, and Mandarin two-talker babble. Although presenting speech in the visual modality alone differs from how speech is typically encountered, this technique has the theoretical benefit of removing a confound of energetic masking that is present in studies assessing target-masker linguistic similarity in the auditory domain alone. Thus, we adopted this methodology because we were primarily interested in testing the strong version of the theory, namely that *linguistic* similarity between a target and masker influences target speech intelligibility. If languages that are perceptually more similar to English provide greater interference than perceptually dissimilar languages, lipreading performance should be better in languages that are more distinct from English, and should be best in quiet or steady-state noise. Thus, we hypothesized that participants would show graded release from masking such that lipreading performance would be most impaired by semantically meaningful English masking, followed by semantically anomalous English masking, Dutch, and Mandarin, and that performance would be best and equivalent in silence and steady-state noise (consistent with the results of Lidestam et al., 2014; Myerson et al., 2016). The semantically meaningful and anomalous conditions were included to assess whether the linguistic content of the masking speech has the potential to differentially mask the target speech. Brouwer and colleagues (2012) showed that meaningful English maskers impaired identification of English target speech more than anomalous maskers (Brouwer et al., 2012), whereas other work has indicated that masking is equivalent for meaningful and anomalous sentences (Calandruccio et al., 2018). In subsequent experiments (2a and 2b) we also test the linguistic similarity hypothesis using auditory targets as well as auditory maskers to more closely attempt to replicate prior work on the linguistic similarity hypothesis. See the Online Supplementary Material for a table

[2] Note that this finding is also in line with the predictions of the changing-state effect (Jones & Macken, 1993): Masking noise that fluctuates impairs performance more than steady-state noise.

summarizing the key differences between the four experiments (Table S1).

Experiments 1a and 1b

Experiment 1a

Method

All data, code for analyses, and stimuli can be accessed via the Open Science Framework at https://osf.io/84zwt/ and the preregistration is available at https://osf.io/jp2fn.

Participants

We collected data from 103 Carleton College undergraduates aged 18–23 years to attain our final pre-registered sample size of 96 participants. Three participants were excluded for technical difficulties during the experiment, and one was excluded for misunderstanding the instructions. Data from the final three participants were discarded because we had pre-registered a sample size of 96 and had complete datasets from 99. For all four experiments reported here, participants had self-reported normal hearing and normal or corrected-to-normal vision, and reported no familiarity with Mandarin or Dutch (the languages of the maskers used), or with Cantonese or German, given their similarity to Mandarin and Dutch. Familiarity was defined as speaking or studying any of the languages, having friends or family who regularly speak any of the languages in front of them, or having lived in a country where any of these languages were spoken. Participants provided written consent before completing the study and were compensated $11 for 1 h of participation. Carleton College's Institutional Review Board approved all research procedures.

Stimuli

Lipreading stimuli The lipreading stimuli were taken from the Build-A-Sentence (BAS) task (see Tye-Murray et al., 2008, 2016). The BAS has previously been demonstrated to have high test-retest reliability, and individual differences in performance on the task are highly correlated with other measures of lipreading (see Feld & Sommers, 2009). Each sentence contained three nouns (e.g., boy, dog, cook) from a closed-set list of 36 words connected by the verb "watched" (e.g., "The *boy* and the *dog* watched the *cook*" or "The *boy* watched the *dog* and the *cook*"; see the accompanying materials on the Open Science Framework for the full word list). The videos were high-resolution files that showed the head and shoulders of a female native English speaker. Each participant was presented with 144 unique sentences for a total of 432 observations per participant (144 sentences × 3 keywords), divided evenly across conditions. Each of the six conditions contained 24 sentences (presented in a randomized order) leading to 72 keywords per condition, and every target word appeared in each condition an equal number of times. Prior work in the auditory domain has used open-set materials (see Calandruccio et al., 2013), but we opted to use the closed-set BAS task to avoid the floor-level performance that often occurs for lipread speech (see Tye-Murray et al., 2010).

Masking stimuli The five auditory masker types were: semantically meaningful English two-talker babble, semantically anomalous (but syntactically normal) English, Dutch, and Mandarin two-talker babble, and speech-shaped noise. All but the Mandarin maskers were identical to those used by Brouwer et al. (2012). Dutch and English anomalous sentences were from the syntactically normal sentence test (see Nye & Gaitenby, 1974; e.g., "The great car met the milk"), and meaningful sentences (English) were from the Harvard/IEEE sentence list (Rothauser et al., 1969; e.g., "Rice is often served in round bowls"). The Mandarin maskers were identical to those used by Van Engen and Bradlow (2007) and Calandruccio et al. (2013), and included sentences that translated to content such as "Your tedious beacon lifted our cab." All masking sentences were produced by two female talkers and were drawn from a pool of 20 sentences (Mandarin) or 100 sentences (all other types of babble).

To create two-talker babble for each condition, individual audio tracks of each talker were equated on total root-mean-square amplitude using Adobe Audition (version 10.1.1.11), and then the two tracks were combined. Some of the tracks of individual talkers used in previous work were not long enough to cover the duration of 24 sentences (the length of each condition). When tracks had to repeat, we offset the two speakers from one another so that within a two-talker babble track, every segment was unique.

To create individual segments of babble, we randomly sampled 4.5 s segments from the two-talker babble of each masker. We matched those segments on LTAS based on Brouwer et al. (2012) to reduce any long-term differences in energetic masking, using MATLAB installed with the Digital Signal Processing Toolbox (The MathWorks, Inc., 2019). First, we matched the babble segments on root-mean-square amplitude and ran a short-time Fourier transform to obtain the spectrum over time for each babble segment, using a 2,048 sample Fourier duration, a 2,048 sample Hamming window, and a 1,024 sample overlap. Next we averaged the spectrum for each babble segment across time and then across all babble segments to obtain the average LTAS. Then we scaled the spectrum of each babble segment to the average LTAS and inverted the short-time Fourier transform to convert each normalized spectrum to a babble segment. Thus, all the individual masker segments across all languages were matched on LTAS

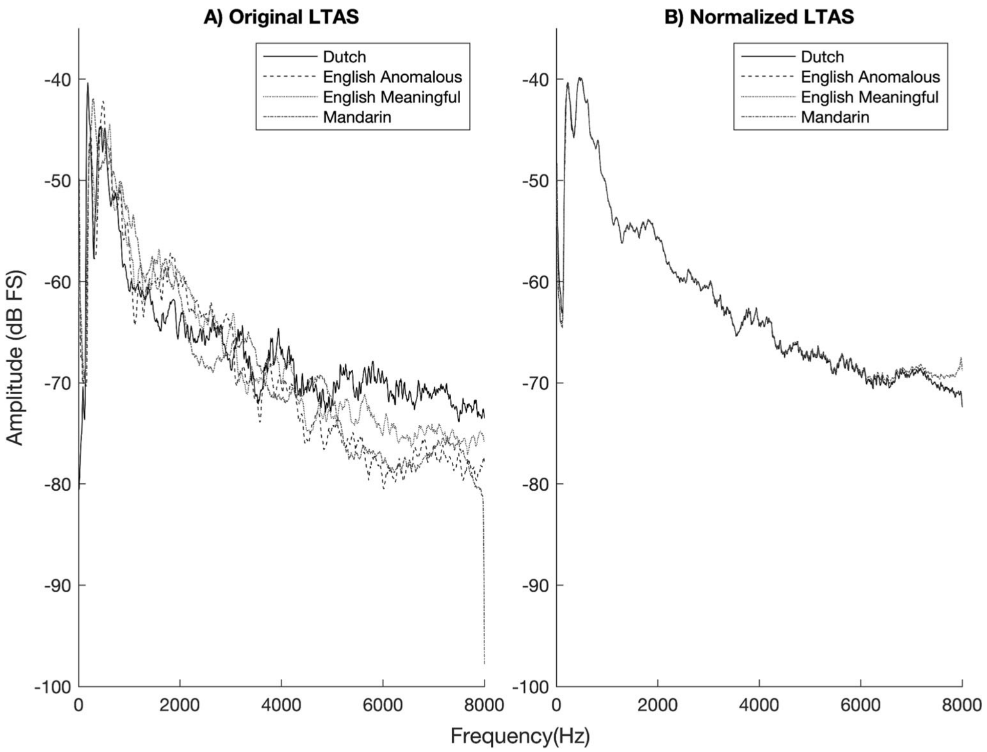

Fig. 1 Long-term average spectrum (LTAS) across all babble tracks for each linguistic masker shown on a log amplitude scale. Disparities in LTAS before normalization (**A**), especially at higher frequencies, are reduced after normalization (**B**). Note that the Mandarin recordings obtained from a previous study on the linguistic similarity hypothesis were recorded at a sampling rate of 16 kHz, which can only accurately represent frequencies up to 8 kHz given Nyquist Sampling (Nyquist, 1928; see also Weisstein, 2022). Both types of English babble and Dutch babble were recorded at 22 kHz

(see Fig. 1). We used this process to generate speech-shaped noise that also had the same LTAS. Note that the original stimuli showed substantial spectral overlap in the frequency range that is typically associated with human speech and only began to diverge markedly above approximately 5 kHz.

Procedure

Testing was conducted on a 21.5 in. iMac computer running SuperLab (Cedrus, version 5.0.5). All auditory stimuli were presented binaurally via Sennheiser HD 280 Pro headphones. In the BAS task, participants watched a video of a sentence to lipread followed by a grid with the 36 keyword choices and repeated aloud what they perceived. Given that the sentence frame always consisted of nouns joined by the verb “watched,” the grid shown after each stimulus only contained the possible target words. In conditions with masking, the noise began 500 ms before the start of the sentence and ended 500 ms after it, at which time the grid with the possible targets was presented, meaning there was no noise during the visual display. Verbal responses were recorded and transcribed offline by research assistants.

Before beginning the experiment, participants were given instructions, shown the 36-item word bank for the BAS lipreading task, and shown one example video in which they were told which sentence to expect. Participants then completed five practice sentences in silence with no feedback, followed by the six experimental blocks corresponding to the six conditions: *English meaningful*, *English anomalous*, *Dutch*, *Mandarin*, *speech-shaped noise*, and *silence*. BAS sentences and maskers were randomly paired such that within a condition, a particular sentence to be lipread appeared with different babble segments across participants. The order of the blocks was counterbalanced across participants according to a balanced Latin Square design, and each target sentence appeared in each babble condition approximately the same number of times across participants.

Results and discussion

Pre-registered analyses

Given the binomial nature of the outcome variable (0 = incorrect, 1 = correct), data were analyzed using generalized linear mixed effects models with a logit link function via the *lme4* package (version 1.1.21; Bates et al., 2014) in R (version 3.5.2; R Core Team, 2020). Nested models were compared via likelihood ratio tests. The fixed effect of interest was masker type (six levels), and we included random intercepts for participants and words and by-participant random slopes for

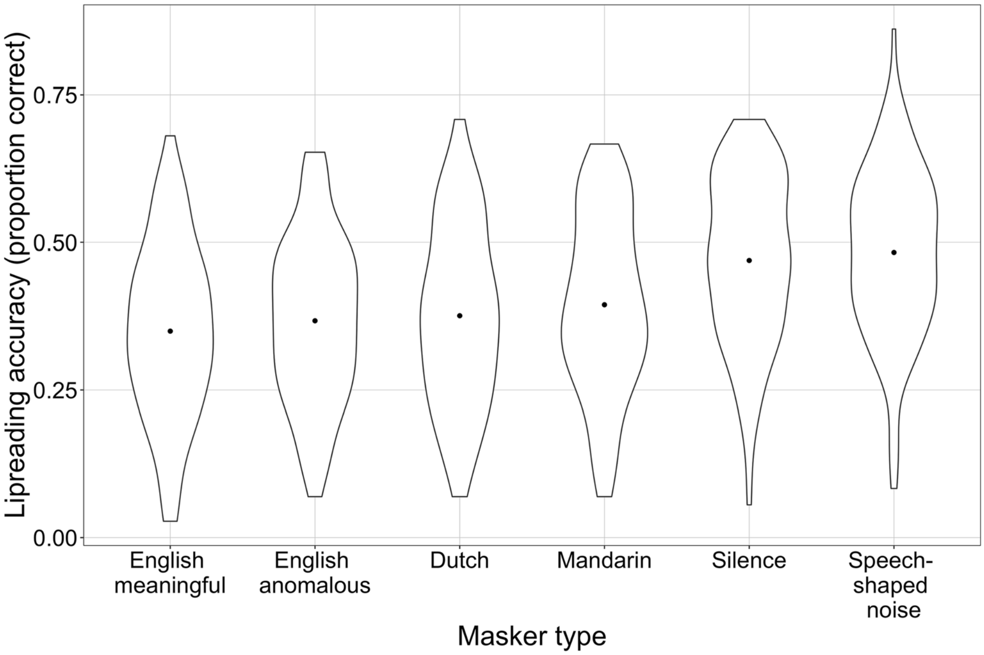

Fig. 2 By-participant average lipreading accuracy in each of the six conditions in Experiment 1a, ordered from lowest to highest accuracy. The dot represents the mean accuracy and the shape represents the distribution of responses for each condition. The accompanying table can be found in the Online Supplementary Material (Table S2)

masker type. We attempted to model the maximal random effects structure justified by the design (following the recommendations of Barr et al., 2013), but models with more complex random effects structures failed to converge (see R script for more details regarding the decisions we made when we encountered convergence and singularity issues). The mean lipreading accuracy was 40.67% and the by-participant standard deviation (i.e., the standard deviation of the participant means) was 15.29%.

To assess whether the type of auditory masker affected lipreading performance, we compared nested models differing only in the fixed effect for masker type. The model that included masker type provided a better fit for the data than the model without it ($\chi^2_5 = 145.85$; $p < .001$), suggesting that the type of background noise influenced lipreading performance. Figure 2 shows by-participant lipreading accuracy in each of the six conditions (i.e., five masker types and silence). The order of lipreading performance in each of the six conditions was numerically consistent with our hypotheses; lipreading accuracy was numerically lowest in the *English meaningful* condition, performance improved as the masker became more linguistically dissimilar from English, and performance was numerically best in the two non-speech conditions (*speech-shaped noise* and *silence*; see also Table S2).

Our next pre-registered set of analyses assessed whether the adjacent conditions, when ordered from lowest to highest lipreading performance, differed significantly from one another. We therefore re-leveled the full model four times (and adjusted *p*-values according to the Holm-Bonferroni method) to obtain the four additional pairwise comparisons of interest: *English anomalous/Dutch*, *Dutch/Mandarin*, *Mandarin/silence*,[3] and *silence/speech-shaped noise* (note that the original model used *English meaningful* as the reference level, so re-leveling was not necessary to obtain the *English meaningful/English anomalous* comparison).

Contrary to our predictions, lipreading performance did not differ for the *English meaningful/English anomalous* (B = -0.11, SE = 0.06, z = -1.85, p = .24), *English anomalous/Dutch* (B = 0.04, SE = 0.06, z = 0.71, p = .48), or *Dutch/Mandarin* (B = -0.11, SE = 0.06, z = -1.88, p = .24) comparisons. However, consistent with our hypothesis, lipreading performance was worse in the *Mandarin* relative to the *silence* condition (B = 0.42, SE = 0.06, z = 7.07, $p < .001$),[4] and did not differ between the *silence* and *speech-shaped noise* conditions (B = 0.07, SE = 0.06, z = 1.32, p = .37). The finding that babble but not speech-shaped noise interfered with lipreading performance relative to silence replicates prior work (Lidestam et al., 2014; Myerson et al., 2016), and suggests that babble provides cognitive interference despite being task-irrelevant.

[3] Our pre-registration specified that we would assess whether the quiet and speech-shaped noise conditions resulted in better lipreading performance than the babble conditions. We therefore conducted the most conservative pairwise comparison to test this hypothesis, which is the *Mandarin/silence* comparison because this comprises the speech and non-speech conditions that resulted in the most similar lipreading performance.

[4] An exploratory analysis revealed that each of the speech maskers impaired lipreading performance relative to the silence condition (*p*s < .001 for all comparisons; see R script for details).

We included the semantically meaningful and anomalous English conditions to test whether linguistic content of the masking speech can differentially interfere with the target speech. The result that anomalous English speech did not differ from meaningful English speech differs from the findings of Brouwer and colleagues, which showed that meaningful speech provided greater interference than anomalous speech when all stimuli were presented in the auditory modality 2012). However, the meaningful and anomalous maskers in that study differed in their prosodic and syntactic features (see Calandruccio et al., 2018), and it may be that the prosody and syntax of the meaningful masker overlapped more with those of the target included in that study than with the ones we used in the current study (note that we used the same maskers as Brouwer and colleagues, but different targets).

One explanation for these findings is that speech maskers are more attentionally salient than non-speech maskers simply as a result of their linguistic nature, and therefore divert attention away from the target speech and impair speech identification to a greater extent. It is also possible that the difference between the speech and non-speech maskers is attributable to the fact that the speech maskers contained acoustic modulation but the steady-state noise did not. Modulating sound can disrupt working memory (i.e., the changing state hypothesis; Jones & Macken, 1993; Salamé & Baddeley, 1989; Tremblay et al., 2001), perhaps because modulation is more attentionally salient than steady-state noise. Thus, the difference in the interference of speech and non-speech maskers may be driven by either acoustic modulation or the linguistic nature of the speech maskers (see Summers & Roberts (2020) for more on the distinction between acoustic-phonetic and linguistic interference in generating informational masking).

The linguistic similarity hypothesis states that more dissimilar languages should provide less informational masking than more similar ones, not necessarily that any step in the linguistic similarity continuum, no matter how small, should result in a detectable change in masking. In other words, our preregistered analyses provided a very strict test of the linguistic similarity hypothesis by only directly comparing languages with the shortest perceptual distances from one another. Thus, we also conducted a set of exploratory analyses to test the linguistic similarity hypothesis with more liberal constraints. These analyses enabled us to test comparisons from prior research conducted in the auditory modality alone such as the difference between anomalous English and Mandarin maskers (Van Engen & Bradlow, 2007) or English meaningful and Dutch maskers (Calandruccio et al., 2013).

Exploratory analyses

Lipreading performance was significantly worse in *English meaningful* relative to *Dutch* ($B = 0.15$, $SE = 0.06$, $z = 2.73$, $p = .006$) and *Mandarin* masking ($B = 0.26$, $SE = 0.06$, $z = 4.81$, $p < .001$), and was significantly worse in *English anomalous* relative to *Mandarin* masking ($B = 0.15$, $SE = 0.05$, $z = 3.13$, $p = .004$; we again adjusted p-values according to the Holm-Bonferroni method for the three values in the exploratory analyses).

Semantically meaningful same-language maskers provided more cross-modal interference than either of the unfamiliar languages tested. These results are in line with the predictions of the linguistic similarity hypothesis and replicate and extend findings of studies that show that masking in another language leads to less interference than masking in a familiar language (e.g., Brouwer et al., 2012). Further, the finding that English anomalous speech provided more masking than Mandarin speech is also consistent with the linguistic similarity hypothesis. However, the strongest evidence for the linguistic similarity hypothesis would be to find differences between two unintelligible maskers that differ in their linguistic similarity to the target speech (i.e., *Dutch* and *Mandarin*). Experiment 1a did not find support for this claim. The aim of Experiment 1b was to assess whether Dutch babble interferes more with English lipreading performance than Mandarin babble does, using a simplified design and a more highly-powered experiment.

Experiment 1b

Method

All data, code for analyses, and materials can be accessed via the OSF at https://osf.io/84zwt/ and the pre-registration is available at https://osf.io/4fyvm.

Experiment 1b followed the conventions of Experiment 1a with several modifications. First, the study included only three masker conditions: *English meaningful*, *Dutch*, and *Mandarin*. We opted to omit the other conditions because the focus was on assessing whether Dutch and Mandarin auditory maskers led to differences in lipread word recognition. Second, the study was conducted online. This change was made to facilitate collecting a large sample of participants while excluding individuals who completed Experiment 1a. Third, because the experiment was conducted online, we instructed participants to type the three keywords of each BAS sentence in a text box rather than repeat them out loud. Finally, the type of background noise was intermixed rather than blocked. We made this decision because we expected that making the masker type less predictable would increase the magnitude of the effect of interest. That is, it may be that participants can more easily “tune out” or habituate to background noise that is consistently presented in the same language, so randomly intermixing the languages may increase the likelihood that more similar languages would be confused, which would be expected to reduce performance in the Dutch but not the Mandarin masking condition. Indeed, previous research has shown that identification of English speech in Dutch babble is worse when these trials are intermixed with English-in-English trials than when they are blocked by background language (Brouwer & Bradlow, 2014).

Participants

A power analysis using an estimated effect size of $d = 0.12$ (the standardized mean difference between the Dutch and Mandarin babble conditions in Experiment 1a) indicated that in order to achieve a power of 0.95, we need a minimum of 124 participants. We opted to more than double this number and analyze data from 250 participants to allay any concerns about additional variability introduced from running the study online. To attain 250 usable datasets, we collected data from 278 participants. Twenty-five participants were excluded based on pre-registered criteria (see below), and the final three participants were not analyzed because the pre-registered sample size had been reached. Washington University in St. Louis' Institutional Review Board approved all research procedures.

We programmed the experiment using Gorilla Experiment Builder (Anwyl-Irvine et al., 2020), and participants were recruited through the Washington University in St. Louis Psychological and Brain Sciences research participant pool via Sona Systems. Data collection occurred between 18 March and 24 April 2020. The experiment took approximately 30 min to complete, and participants were given 0.5 credits for their time (in accordance with departmental policies for use of the subject pool). Due to constraints with presentation of certain file types online, participants could only complete the experiment using Google Chrome.

Exclusion criteria To confirm that participants attended to the task and did not turn off their audio to avoid listening to the background noise, we added periodic auditory attention checks during the experiment. For nine additional BAS trials, instead of playing the background babble, participants were presented with a sentence by a single talker that said "If you can hear this, type ___." The words that completed the sentences were chosen to be intelligible and unique from the words used in the BAS task. Participants were told during the instructions that if they heard such a sentence, they should type the final word rather than trying to lipread the sentence. These sentences were presented at a level 25 dB quieter than the babble to ensure that if participants removed their headphones or turned down the volume they would be unable to complete the task. Participants who missed more than one-third of the attention check keywords were excluded from the analysis (N = 9).

After completing the experiment, participants were asked if they turned the volume down or off to avoid listening to the background noise, or if they did in fact have experience with Dutch, German, Mandarin, or Cantonese. Participants were told that this would not affect their credit for completing the experiment, and were urged to respond honestly. An additional 18 participants were excluded based on the results of these questions.

Stimuli

Lipreading stimuli Participants saw 72 unique BAS sentences (a subset of those used in Experiment 1a) with three keywords each, resulting in 216 words for each participant (24 sentences or 72 words per condition).

Masking stimuli Masking stimuli were *English meaningful*, *Dutch*, and *Mandarin* from Experiment 1a.

Procedure

Participants were first presented with a sentence at a level 25 dB below the level at which the background babble would be presented, and were instructed to wear headphones and adjust their computer volume so they could hear the sentence at a very quiet level. The purpose of this volume adjustment phase was to ensure that participants could hear the auditory attention check sentences. After setting their volume, participants completed a headphone screening for web-based auditory experiments (Woods et al., 2017). In this task, participants were presented with three 200-Hz tones—one of which was 180° out of phase across stereo channels—on each of six trials, and after the third tone participants indicated which of the three tones was the quietest. Due to phase cancellation, the amplitude of the unique tone is difficult to distinguish from that of the other two tones when the listener is not using headphones, but can be readily distinguished when the listener is using headphones. If the participant passed the headphone screening, they then completed six BAS practice trials without background noise. If they did not pass the screening, they were informed that the experiment was over, but were allowed to rebook the experiment if they wanted to try again with headphones.

The procedure for the BAS trials was similar to the procedure in Experiment 1a, with a few alterations. After each BAS trial, participants were concurrently presented with a word bank and a text box and instructed to type the three keywords into the text box, separated by spaces, then press enter. After indicating that they had responded, the screen went blank for 100 ms, followed by a fixation cross in the center of the screen for 1,500 ms and another 100 ms blank screen. Then, the next BAS trial began playing. Following all of the BAS trials, participants completed the honesty check questionnaire described above. Unlike in Experiment 1a—in which videos and babble tracks were paired randomly—target videos in this experiment were yoked to babble tracks, but given the construction of the BAS, every word still appeared in every condition an equal number of times for every participant.

An experimenter checked all typed responses, allowing pluralizations, extraneous punctuation, and homophones to be counted as correct. Responses with typos were counted as correct only when the error was a single-letter addition or deletion. Responses that were one letter away from the target but resulted in a different word were not counted as correct (e.g., fog/frog; mouse/moose).

Results and discussion

Pre-registered analyses

Data analysis followed the conventions of Experiment 1a, but the fixed effect of interest had three levels rather than six. The final model included random intercepts for participants and words as well as by-participant and by-item random slopes for masker type. The mean lipreading accuracy was 40.37% and the by-participant standard deviation was 18.04%.

To assess whether the type of background noise influenced lipreading performance, we compared nested models differing only in the presence of the fixed effect for masker type. A likelihood ratio test indicated that masker type significantly influenced lipreading performance ($\chi^2_2 = 14.26$; $p < .001$). As in Experiment 1a, the relative ordering of lipreading performance in each of the three conditions was consistent with our hypothesis; performance was numerically worst in *English* babble and best in *Mandarin* babble, with *Dutch* babble resulting in numerically slightly poorer performance than *Mandarin* babble (see Fig. 3 and Table S3).

Consistent with our predictions and the results of Experiment 1a, lipreading performance was significantly worse in *English* than both *Dutch* (B = -0.35, SE = 0.09, z = -3.93, $p < .001$) and *Mandarin* (B = -0.38, SE = 0.11, z = -3.58, $p < .001$) babble. However, the difference in lipreading performance between *Dutch* and *Mandarin* babble was not significant (B = 0.03, SE = 0.09, z = 0.38, p = .71). Thus, the presence of English maskers led to poorer lipreading performance for English target speech than the Dutch and Mandarin maskers did, but the linguistic similarity between the unintelligible maskers and the target speech did not differentially affect lipreading performance.

Exploratory analysis

Finally, to assess whether the hypothesized effects of greater interference from Dutch than Mandarin maskers might emerge when the sample size was even larger, we combined the relevant conditions from Experiment 1a (i.e., data from the Dutch, Mandarin, and English meaningful conditions) with

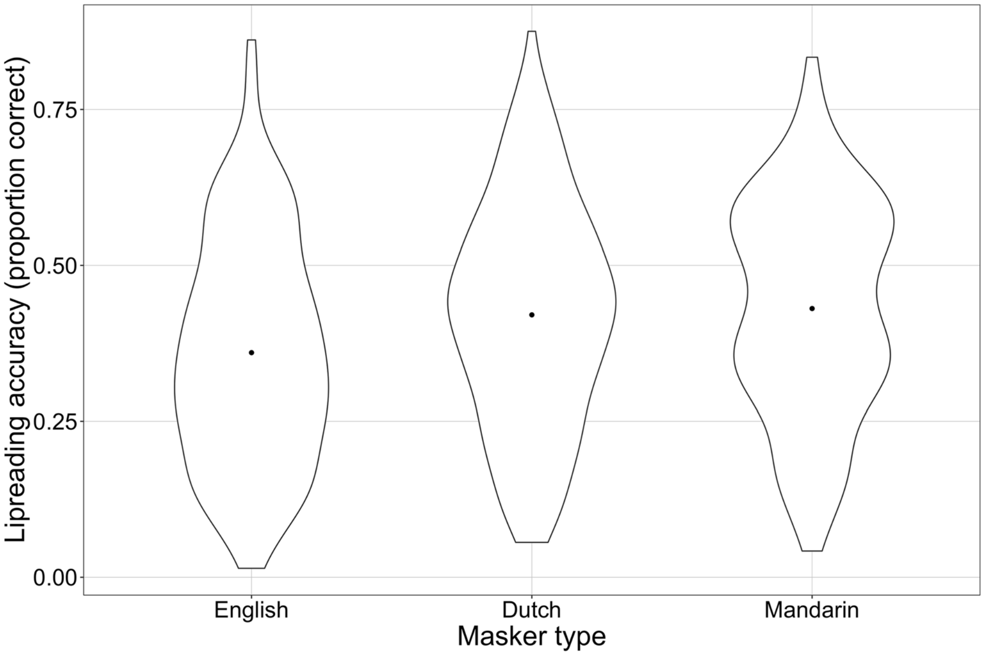

Fig. 3 By-participant mean lipreading accuracy in each of the three conditions in Experiment 1b, ordered from lowest to highest accuracy. The dot represents the mean accuracy and the shape represents the distribution of responses for each condition. The accompanying table can be found in the Online Supplementary Material (Table S3)

the data from Experiment 1b. This analysis, which included data from 346 participants, also revealed no difference in lipreading accuracy between the Dutch and Mandarin conditions (B = 0.05, SE = 0.07, z = 0.76, p = .45). Although the differences in the Dutch and Mandarin masking conditions were in the anticipated direction, the difference is not statistically reliable, even with a very large sample.

Experiment 1: Discussion

Across two experiments with a large combined sample size, we did not find any evidence that the linguistic similarity of the masking language to English affects visual-only word recognition accuracy. The results showed that lipreading English speech was impaired more by the presence of English two-talker babble than two-talker babble in either Mandarin or Dutch, and that all types of speech babble impaired lipreading more than speech-shaped masking noise. Research on the linguistic similarity hypothesis has always involved auditory maskers and speech, and makes no explicit claims about cross-modal speech. However, if the interference from the masker is caused by cognitive rather than acoustic interference, an unintelligible masker that is more similar to the target speech (Dutch) should interfere more with speech processing than one that is less similar (Mandarin), regardless of modality. We did not find support for this claim using visual speech targets and auditory maskers.

If Dutch had provided more interference than Mandarin, it would have provided cross-modal support for the prediction of the linguistic similarity hypothesis that greater linguistic similarity leads to greater interference. However, the null effect reported here—not finding a difference in interference between the Dutch and Mandarin maskers—does not necessarily provide evidence against the linguistic similarity hypothesis. That is, the effects of graded similarity may be expected to emerge for auditory maskers with auditory targets, but fail to emerge in cross-modal situations because of differences in how attentional or memory resources are allocated cross-modally, or reduced demands of stream segregation when the target and masker are in separate modalities.

Although the current work does not definitively challenge the linguistic similarity hypothesis, a close examination of the literature shows that prior work using auditory targets and auditory maskers has also not found strong support for graded interference. The linguistic similarity hypothesis states that "the more similar the target and the masker speech, the harder it is to segregate the two streams effectively" (Brouwer et al., 2012, p. 1449). This definition implies that it should be possible to select maskers that differ from the target language along a continuum of similarity and demonstrate graded interference from those maskers. In practice, however, the majority of research on the linguistic similarity hypothesis has compared maskers that are either the same language or a different language from the target speech (e.g., Brouwer et al., 2012; Van Engen & Bradlow, 2007). The findings from previous work, like those reported here, are in line with the predictions of the hypothesis: Maskers in the same language as the target speech interfere more than maskers in different languages. However, this approach evaluates the influence of linguistic *sameness* rather than linguistic *similarity*. This distinction may be subtle but is important; if these effects only occur when the target and masker are presented in the same language, the interference may be driven by intelligibility rather than linguistic similarity, per se.[5]

Thus, to show graded interference, as the linguistic similarity hypothesis predicts, it is necessary to include multiple maskers that vary in their similarity to the target speech. However, the only study that has included maskers in multiple languages (English, Dutch, and Mandarin; Calandruccio et al., 2013) did not match the maskers on LTAS. Further, they found that although Dutch speech masked more than Mandarin speech—consistent with the predictions of the linguistic similarity hypothesis—speech-shaped noise that was generated based on the spectral characteristics of the Dutch speech also provided more masking than speech-shaped noise generated from the Mandarin speech, suggesting that those findings were due in part to energetic rather than informational masking. Therefore, in Experiments 2a and 2b, we again used LTAS-matched English, Dutch, and Mandarin maskers but used *auditory* English targets.

Experiments 2a and 2b

Experiment 2a

Method

All data, code for analyses, and materials can be accessed via the OSF at https://osf.io/84zwt/ and the pre-registration is available at https://osf.io/vw7r2.

[5] There is some evidence that intelligibility alone cannot account for findings supporting the linguistic similarity hypothesis. That study showed that Dutch-English bilinguals, like monolingual English speakers, receive greater interference from English relative to Dutch speech when the target language is English (Brouwer et al., 2012). If the linguistic similarity hypothesis were driven only by the intelligibility of the masker, then Dutch-English bilinguals should receive equivalent interference from Dutch and English, so the intelligibility of the masker is not the only factor contributing to speech-on-speech masking. However, Dutch-English bilinguals receive greater interference from Dutch when recognizing English speech than do monolingual English speakers, indicating that the intelligibility of the masker clearly plays an important role in speech-on-speech masking (see also Calandruccio & Zhou, 2014; Van Engen, 2010).

Participants

A power analysis using an estimated effect size of $d = 0.3$ and a power of 0.95 showed that we would need a minimum of 101 participants. The smallest effect size of interest from Calandruccio et al. (2013) was 1.08, which would result in a sample size of 7, which we decided was too small. The effect size difference between English and Dutch babble on lipreading performance in Experiment 1a was 0.35, so running a power analysis with $d = 0.3$ gave us a conservative estimate, which would allow us to detect smaller effects of interest. We opted to approximately double this sample size of 101 and collect data from 204 participants to enable a balanced design.

The experiment was programmed and presented via Gorilla Experiment Builder (Anwyl-Irvine et al., 2020), and participants were recruited through Prolific (prolific.co). Participants' ages ranged from 18 to 41 years (M = 26.59, SD = 5.12). Data collection began on 26 February 2021 and ended on 3 May 2021. The experiment took approximately 20 min to complete, and participants were compensated $4.00 for their time. In order to reach our pre-registered sample size of 204, a total of 230 participants completed Experiment 2a. The Carleton College Institutional Review Board approved the procedures for Experiments 2a and 2b.

Exclusion criteria After the experiment, participants were asked if they had knowledge of Mandarin, Cantonese, Dutch, or German, and if they paid attention to the experiment to the best of their abilities. They were instructed that their answers would not affect their compensation. If participants indicated that they had knowledge of the languages listed or that they did not pay attention, they were excluded from the main analyses (N = 16).

To ensure that participants were paying attention to the task, we included nine auditory-only attention check sentences spoken by a single speaker following the format of "If you can hear this, type ___." This audio clip was played in place of the background noise, and the target stimuli were muted. These sentences were played at a level 10 dB quieter than the level of the speech stimuli. If participants responded incorrectly to more than three of these sentences, we excluded them from the main analyses (N = 4). Finally, we excluded participants if their speech identification accuracy was worse than three standard deviations below the mean in any condition (N = 6).

Stimuli

Target stimuli Following the procedure of Calandruccio et al. (2013), the target stimuli consisted of Bamford-Kowal-Bench (BKB) sentences (Bench et al., 1979). Ninety-six sentences were obtained from an existing source (Van Engen, 2010), produced by a female native speaker of American English with no obvious regional accent. Each sentence had three keywords for scoring (i.e., "The *match fell* on the *floor*"), resulting in a total of 288 words for each participant.

Masking stimuli The maskers were identical to those in Experiment 1b (all female speakers), and the signal-to-noise ratio (SNR) was set to -7 dB. We also included a condition without masking noise (*silence*) to match the procedures of Calandruccio et al. (2013) and assess whether even the least detrimental masker still impaired performance relative to a condition with no masker. Stimuli were counterbalanced across masking conditions such that across participants, all sentences appeared in all conditions.

Procedure

Before being directed to Gorilla, participants were screened on Prolific to ensure that they were currently located in the USA, their first language was English, and they were between the ages of 18 and 35 years (note that one participant reported being 41 and one reported being 36 years of age despite this restriction). Participants were then presented with five practice BKB sentences in silence and were instructed to type what they heard after the sentence played. Following each practice sentence, the correct answer was displayed on the screen. The main experiment began following completion of the five practice trials.

Between each trial, a fixation cross appeared on the center of the screen for 500–2,000 ms (varying in increments of 500 ms). The fixation cross stayed on the screen as auditory stimuli played, and immediately following the trial a text box appeared with the label "Type what you heard." The nine attention check sentences were intermixed with critical trials.

The accuracy of all trials was coded first by Autoscore (Borrie et al., 2019), an automated R package for assessing the accuracy of typed input that counts predetermined typos or grammar mistakes as correct, and then manually checked by an experimenter. Extraneous punctuation was eliminated, and homophones (e.g., pair/pear) were counted as correct. Typos were counted as correct only when the error was a single-letter addition or deletion from a correct word (e.g., bear/ber), the input was one keystroke away from the correct answer (e.g., tree/gree), or the input was a common misspelling of the target word (e.g., exercise/exersise), provided that the participant's answer did not result in a different word (e.g., fog/frog). Pluralizations were not counted as correct (e.g., dog/dogs).

Results and discussion

Data analysis followed the conventions of the previous studies. The final model included random intercepts for participants and words as well as by-participant and by-item random slopes for masker type (i.e., the maximal random effects structure justified by the design; Barr et al., 2013).

Consistent with our predictions, all maskers impaired performance relative to the silence condition: *English* (B = -6.76, SE = 0.25, z = -26.98, p < .001), *Dutch* (B = -5.79, SE = 0.19, z = -30.41, p < .001), and *Mandarin* (B = -6.25, SE = 0.20, z = -31.85, p < .001). In addition, performance was significantly worse in *English* than in both *Dutch* (B = 0.97, SE = 0.21, z = 4.66, p < .001) and *Mandarin* (B = 0.51, SE = 0.22, z = 2.36, p = .02) babble. In contrast to the results of the previous experiments and the predictions of the linguistic similarity hypothesis, word identification was significantly worse in *Mandarin* than *Dutch* babble (B = -0.46, SE = 0.18, z = -2.56, p = .02; see Fig. 4 and Table S4). All p-values were adjusted according to the Holm-Bonferroni method.

The fact that *Mandarin* impaired performance more than *Dutch* is unexpected and difficult to account for, especially considering that Calandruccio et al. (2013) showed better performance in the Mandarin masker than the Dutch masker—in line with the predictions of the linguistic similarity hypothesis—using similar stimuli. One explanation for the discrepancy between our results and those of Calandruccio and colleagues could be that the effects found in their study were driven by acoustical differences in the maskers, and the LTAS matching we employed reduced those differences. Another difference between Experiment 2a and Calandruccio and colleagues' study was that word identification accuracy in our study was lower overall, perhaps as a result of running the study online and therefore delivering auditory stimuli in a less controlled way. If the effects of linguistic similarity are dependent upon the difficulty of the task, then we may not have detected them here because accuracy was poor. We therefore conducted an additional experiment with an easier SNR to attempt to match performance to that in Calandruccio et al. (2013).

Experiment 2b

Method

All data, code for analyses, and materials can be accessed via the OSF at https://osf.io/84zwt/ and the pre-registration is available at https://osf.io/xyj8m.

Participants

Based on the results from the first iteration of this experiment, we opted to collect data from 100 participants. Any differences that only appear when N > 100 are deemed too small for our interest. A power analysis based on the observed effect size (d = 1.08) from Calandruccio et al. (2013) revealed that we would need at least seven participants to achieve a power of .95.

Participants ranged from 18 to 45 years of age (M = 28.37, SD = 4.87; note that one participant reported being 45 years of age despite the 18–35 age restriction we set on Prolific). Data collection began on 4 May 2021 and ended on 7 May 2021.

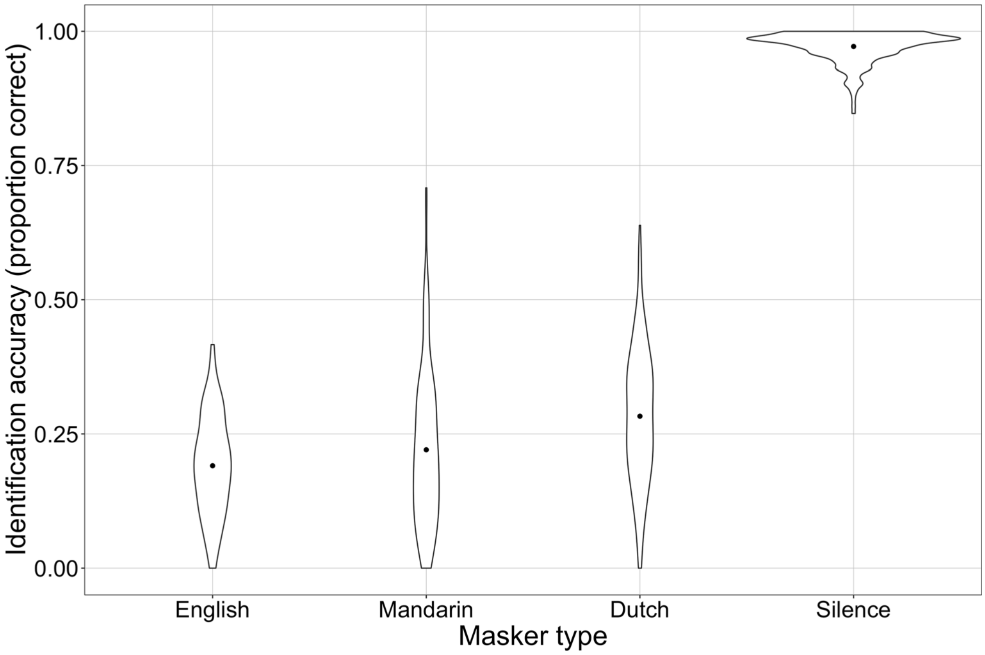

Fig. 4 By-participant mean sentence identification accuracy in each of the four conditions in Experiment 2a, ordered from lowest to highest accuracy. The dot represents the mean accuracy and the shape represents the distribution of responses for each condition. The accompanying table can be found in the Online Supplementary Material (Table S4)

The experiment took approximately 20 min to complete, and participants were compensated $4.00 for their time. To reach our pre-registered sample size of 100, we ran 115 participants.

Exclusion criteria Exclusion criteria were identical to those in Experiment 2a. Eleven participants were excluded for having knowledge of Dutch, German, Cantonese, or Mandarin; one was excluded for failing the auditory attention check; and one was excluded for having poor speech identification accuracy (worse than three standard deviations below the mean). Given that this resulted in two extra participants in one of our counterbalanced orders, we removed the final two participants who finished that order, resulting in a sample size of 100 participants.

Stimuli

Target stimuli Stimuli were identical to those in Experiment 2a.

Masking stimuli The masking stimuli were identical to those in Experiment 2a, but the SNR was changed to -4 dB to more closely match performance to that in Calandruccio et al. (2013).

Procedure

The procedure for Experiment 2b was identical to that of Experiment 2a.

Results and discussion

Changing the SNR was successful in improving performance for the masked speech. Word identification accuracy for the English masker condition in Calandruccio et al. (2013) was approximately 37%, and here was 37.56%. Thus, any deviations in our results from theirs are not likely to be attributable to differences in the difficulty of the task.

All models included random intercepts for participants and items, as well as by-participant and by-item random slopes for masker type. As in Experiment 2a, all maskers impaired performance relative to the silence condition: *English* (B = -5.41, SE = 0.28, z = -19.19, p < .001), *Dutch* (B = -4.74, SE = 0.26, z = -18.34, p < .001), and *Mandarin* (B = -4.62, SE = 0.25, z = -18.16, p < .001; all p-values were again adjusted via the Holm-Bonferroni method). In addition, performance was significantly worse in *English* than in both *Dutch* (B = 0.66, SE = 0.20, z = 3.37, p < .001) and *Mandarin* (B = 0.79, SE = 0.20, z = 3.99, p < .001) babble. However, word identification did not differ between *Mandarin* and *Dutch* babble (B = 0.12, SE = 0.17, z = 0.72, p = .47; see Fig. 5 and Table S5).

Experiment 2b provides the closest approximation to a direct replication of Calandruccio et al. (2013), with the key difference of matching the masking stimuli on LTAS. Despite using very similar methods and matching the level of difficulty for the English maskers, we did not find support for the finding that Dutch maskers provide more interference than Mandarin maskers for English target speech.

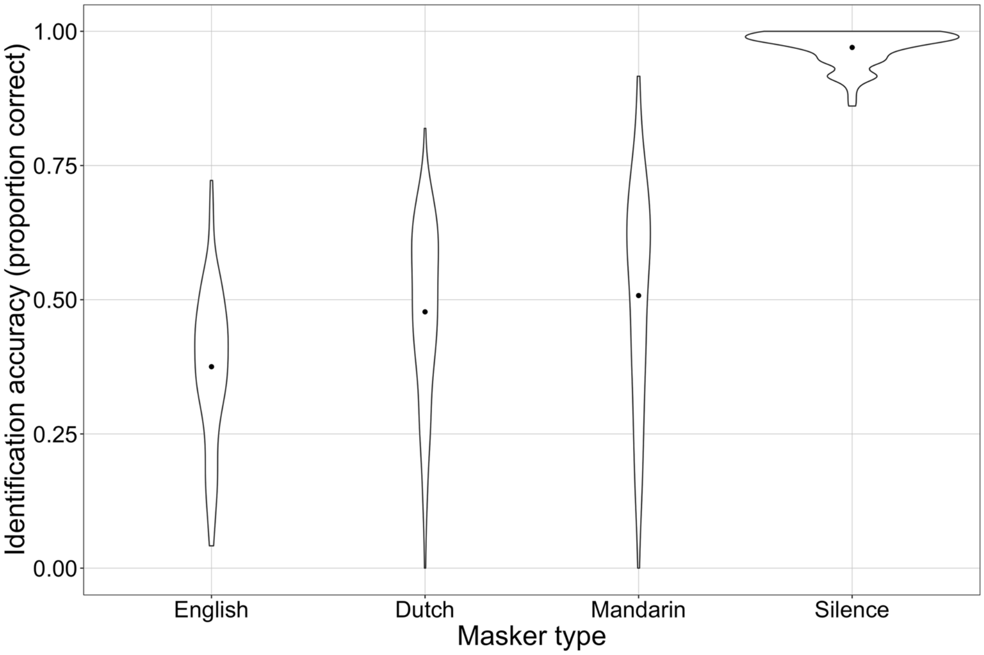

Fig. 5 By-participant mean sentence identification accuracy in each of the four conditions in Experiment 2b, ordered from lowest to highest accuracy. The dot represents the mean accuracy and the shape represents the distribution of responses for each condition. The accompanying table can be found in the Online Supplementary Material (Table S5)

General discussion

The linguistic similarity hypothesis predicts that maskers that are more linguistically similar to the target speech should impair identification accuracy more than maskers that are less similar. Across four experiments—including two modalities of target speech, in-lab and online samples, two SNRs, and a total of 650 participants—we did not find any evidence that a more linguistically similar masker (Dutch) impairs performance to a greater extent than a less linguistically similar masker (Mandarin). However, we *did* find robust evidence that a same-language masker (English) resulted in poorer performance than different-language maskers (Dutch and Mandarin). Indeed, in all four experiments, the English masking speech resulted in poorer levels of performance than maskers produced in any other language. One explanation for the finding that maskers produced in the same language as the target impair intelligibility more than maskers produced in other languages is that the target and masker streams are more difficult to segregate when they are produced in the same language (Brouwer et al. 2012).[6] That is, listeners may ascribe features of the masker stream to the target stream, thereby impairing intelligibility of the target.

Alternatively or in addition, these findings may be attentional in nature Same-language maskers may be more attentionally demanding than maskers produced in unfamiliar languages. This is consistent with prior work on the irrelevant speech effect showing that performance on visual working memory tasks is more impaired by masking speech in a listener's native language than in an unfamiliar language (Ellermeier et al., 2015). This could also explain why intelligible maskers provide the most interference—intelligible words or phrases in the masker stream may capture attention, thereby drawing attention away from the target stream so listeners fail to notice information in that stream (see Van Engen, 2010).

Even if the masker language is unintelligible, overlapping phonology may lead to lexical activation in the target language (see Spivey & Marian, 1999), which may be expected to capture attention as well. Although unintelligible maskers are unlikely to lead to lexical intrusions (except in the case of cognates), they may produce interference before lexical objects have been formed (i.e., "acoustic-phonetic interference"; Summers & Roberts, 2020). If overlapping phonology leads to lexical activation in the target language, and Dutch has more overlapping phonology with English than Mandarin does, why then does Dutch not impair identification of English speech to a greater extent than Mandarin? Listeners are extraordinarily sensitive to sub-phonemic cues and may therefore be able to "tune out" maskers produced in different languages, even if those languages are linguistically similar to the target speech. Indeed, Spanish-English bilinguals demonstrate fine-grained sensitivity to allophonic variation in the form of differences in the voice onset time between the same phoneme produced in Spanish versus English (i.e., sensitivity to the fact that the /p/ in the Spanish word "playa" has a shorter voice onset time than the /p/ in the English word "pliers"; Ju & Luce, 2004).

The current work also provides support for the efficacy of cross-modal masking paradigms. The auditory-only conditions (Experiments 2a and 2b) provided more direct tests of the linguistic similarity hypothesis because the hypothesis typically refers to auditory processing, but the cross-modal masking conditions (Experiments 1a and 1b) yielded the same pattern of results: All speech maskers impaired word recognition accuracy relative to silence, but the same-language masker impaired it the most.

Thus, although there is substantial support for the claim that maskers produced in the same language as the target provide greater interference than those produced in a different language, we found no evidence that linguistic similarity between the target and two unfamiliar languages varying in their degree of similarity to the target (e.g., Dutch and Mandarin for monolingual English speakers) affects target speech identification. Previous results that have seemed to contradict this claim may have been driven by spectral rather than linguistic differences among maskers (see Calandruccio et al., 2013). In the absence of evidence that unintelligible maskers that are similar to the target impair speech recognition more than dissimilar ones, the linguistic similarity hypothesis should be refined to reflect the existing evidence—namely that there is more release from masking when the masker language differs from the target speech than when it is the same as the target speech.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13414-022-02486-3.

Acknowledgements This work was supported by Carleton College, Washington University in St. Louis, the National Science Foundation through a Graduate Research Fellowship awarded to Violet Brown (DGE-1745038), and the National Institutes of Health (National Institute on Deafness and Communication Disorders) via a grant to Julia Strand (R15-DC018114). We wish to thank the Carleton Perception Lab research assistants for assistance with data collection and fruitful discussions. We are grateful to Susanne Brouwer for sharing speech stimuli and the code to perform the LTAS matching procedure and to Brent Spehar for sharing the Build-A-Sentence stimuli.

[6] Note, however, that this explanation is only pertinent to conditions in which the target and masker are presented in the same modality; stream segregation should be a relatively trivial task when the target is visual and the masker is auditory.

References

Agus, T. R., Akeroyd, M. A., Gatehouse, S., & Warden, D. (2009). Informational masking in young and elderly listeners for speech masked by simultaneous speech and noise. *The Journal of the Acoustical Society of America, 126*(4), 1926–1940.

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.

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

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., & Green, P. (2014). *Package "lme4"* (Version 1.1-15). R foundation for statistical computing, Vienna, 12. https://github.com/lme4/lme4/

Bench, J., Kowal, A., & Bamford, J. (1979). The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *British Journal of Audiology, 13*(3), 108–112.

Borrie, S. A., Barrett, T. S., & Yoho, S. E. (2019). Autoscore: An open-source automated tool for scoring listener perception of speech. *The Journal of the Acoustical Society of America, 145*(1), 392.

Bradlow, A., Clopper, C., Smiljanic, R., & Walter, M. A. (2010). A perceptual phonetic similarity space for languages: Evidence from five native language listener groups. *Speech Communication, 52*(11-12), 930–942.

Bregman, A. S., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology, 89*(2), 244–249.

Brouwer, S. (2017). Masking release effects of a standard and a regional linguistic variety. *The Journal of the Acoustical Society of America, 142*(2), EL237.

Brouwer, S. (2019). The role of foreign accent and short-term exposure in speech-in-speech recognition. *Attention, Perception & Psychophysics, 81*(6), 2053–2062.

Brouwer, S., & Bradlow, A. R. (2014). Contextual variability during speech-in-speech recognition. *The Journal of the Acoustical Society of America, 136*(1), EL26–EL32.

Brouwer, S., Van Engen, K. J., Calandruccio, L., & Bradlow, A. R. (2012). Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *The Journal of the Acoustical Society of America, 131*(2), 1449–1464.

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

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C., Kiessling, J., Kotby, M. N., Nasser, N. H. A., El Kholy, W. A. H., Nakanishi, Y., Oyer, H., Powell, R., Stephens, D., Meredith, R., & Ludvigsen, C. (1994). An international comparison of long-term average speech spectra. *The Journal of the Acoustical Society of America, 96*(4), 2108–2120.

Calandruccio, L., Brouwer, S., Van Engen, K. J., Dhar, S., & Bradlow, A. R. (2013). Masking release due to linguistic and phonetic dissimilarity between the target and masker speech. *American Journal of Audiology, 22*(1), 157–164.

Calandruccio, L., Buss, E., Bencheck, P., & Jett, B. (2018). Does the semantic content or syntactic regularity of masker speech affect speech-on-speech recognition? *The Journal of the Acoustical Society of America, 144*(6), 3289.

Calandruccio, L., Dhar, S., & Bradlow, A. R. (2010). Speech-on-speech masking with variable access to the linguistic content of the masker speech. *The Journal of the Acoustical Society of America, 128*(2), 860–869.

Calandruccio, L., Wasiuk, P. A., Buss, E., Leibold, L. J., Kong, J., Holmes, A., & Oleson, J. (2019). The effect of target/masker fundamental frequency contour similarity on masked-speech recognition. *The Journal of the Acoustical Society of America, 146*(2), 1065.

Calandruccio, L., & Zhou, H. (2014). Increase in speech recognition due to linguistic mismatch between target and masker speech: Monolingual and simultaneous bilingual performance. *Journal of Speech, Language, and Hearing Research: JSLHR*. https://doi.org/10.1044/2013_JSLHR-H-12-0378

Campbell, T., Beaman, C. P., & Berry, D. C. (2002). Changing-state disruption of lip-reading by irrelevant sound in perceptual and memory tasks. *The European Journal of Cognitive Psychology, 14*(4), 461–474.

Carhart, R., Tillman, T. W., & Greetis, E. S. (1969). Perceptual masking in multiple sound backgrounds. *The Journal of the Acoustical Society of America, 45*(3), 694–703.

Dannenbring, G. L., & Bregman, A. S. (1978). Streaming vs. fusion of sinusoidal components of complex tones. *Perception & Psychophysics, 24*(4), 369–376.

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., & Jr., G. K. (2003). Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. *The Journal of the Acoustical Society of America, 114*(1), 368–379.

Ellermeier, W., Kattner, F., Ueda, K., Doumoto, K., & Nakajima, Y. (2015). Memory disruption by irrelevant noise-vocoded speech: Effects of native language and the number of frequency bands. *The Journal of the Acoustical Society of America, 138*(3), 1561–1569.

Feld, J. E., & Sommers, M. S. (2009). Lipreading, processing speed, and working memory in younger and older adults. *Journal of Speech, Language, and Hearing Research: JSLHR, 52*, 1555–1565.

Festen, J. M., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America, 88*(4), 1725–1736.

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

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

Garcia Lecumberri, M. L., & Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. *The Journal of the Acoustical Society of America, 119*(4), 2445–2454.

Helfer, K. S., Chevalier, J., & Freyman, R. L. (2010). Aging, spatial cues, and single- versus dual-task performance in competing speech perception. *The Journal of the Acoustical Society of America, 128*(6), 3625–3633.

Ihlefeld, A., & Shinn-Cunningham, B. (2008). Spatial release from energetic and informational masking in a selective speech identification task. *The Journal of the Acoustical Society of America, 123*(6), 4369–4379.

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science, 24*(10), 1995–2004.

Jones, D. M., & Macken, W. J. (1993). Irrelevant tones produce an irrelevant speech effect: Implications for phonological coding in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(2), 369–381. https://doi.org/10.1037//0278-7393.19.2.369

Ju, M., & Luce, P. A. (2004). Falling on sensitive ears: Constraints on bilingual lexical activation. *Psychological Science, 15*(5), 314–318.

Kidd, G., Mason, C. R., & Gallun, F. J. (2005). Combining energetic and informational masking for speech identification. *The Journal of the Acoustical Society of America, 118*(2), 982–992.

Kidd, G., Mason, C. R., Rohtla, T. L., & Deliwala, P. S. (1998). Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns. *The Journal of the Acoustical Society of America, 104*(1), 422–431.

Lidestam, B., Holgersson, J., & Moradi, S. (2014). Comparison of informational vs. energetic masking effects on speechreading performance. *Frontiers in Psychology, 5*, 639.

Myerson, J., Spehar, B., Tye-Murray, N., Van Engen, K., Hale, S., & Sommers, M. S. (2016). Cross-modal informational masking of lipreading by babble. *Attention, Perception & Psychophysics, 78*(1), 346–354.

Newman, R. S., & Evers, S. (2007). The effect of talker familiarity on stream segregation. *Journal of Phonetics, 35*(1), 85–103.

Nye, P. W., & Gaitenby, J. H. (1974). The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. *Haskins Laboratories Status Report on Speech Research, 38*(169–190), 43.

Nyquist, H. (1928). Certain Topics in Telegraph Transmission Theory. *Transactions of the American Institute of Electrical Engineers, 47*(2), 617–644.

R Core Team. (2020). R 4.0.2. R Foundation for Statistical Computing Vienna, Austria.

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science, 17*(6), 405–409.

Rothauser, E. H., Chapman, W. D., Guttman, N., Silbiger, H. R., Hecker, M. H. L., Urbanek, G. E., Nordby, K. S., & Weinstock, M. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics, 17*(3), 225–246.

Rothpletz, A. M., Wightman, F. L., & Kistler, D. J. (2012). Informational masking and spatial hearing in listeners with and without unilateral hearing loss. *Journal of Speech, Language, and Hearing Research: JSLHR, 55*(2), 511–531.

Salamé, P., & Baddeley, A. (1989). Effects of Background Music on Phonological Short-Term Memory., *41*(1), 107–122. The Quarterly Journal of Experimental Psychology Section A. https://doi.org/10.1080/14640748908402355

Spivey, M. J., & Marian, V. (1999). Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science, 10*(3), 281–284.

Strand, J. F., & Sommers, M. S. (2011). Sizing up the competition: Quantifying the influence of the mental lexicon on auditory and visual spoken word recognition. *The Journal of the Acoustical Society of America, 130*(3), 1663–1672.

Summers, R. J., & Roberts, B. (2020). Informational masking of speech by acoustically similar intelligible and unintelligible interferers. *The Journal of the Acoustical Society of America, 147*(2), 1113.

The MathWorks, Inc. *(2019). MATLAB and Statistics Toolbox (Release 2019a) [Computer software*.

Tremblay, S., MacKen, W. J., & Jones, D. M. (2001). The impact of broadband noise on serial memory: Changes in band-pass frequency increase disruption. *Memory, 9*(4), 323–331.

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

Tye-Murray, N., Sommers, M. S., Spehar, B., Myerson, J., Hale, S., & Rose, N. S. (2008). Auditory-visual discourse comprehension by older and young adults in favorable and unfavorable conditions. *International Journal of Audiology, 47*(sup2), S31–S37.

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

Van Engen, K. J. (2010). Similarity and familiarity: Second language sentence recognition in first- and second-language multi-talker babble. *Speech Communication, 52*(11-12), 943–953.

Van Engen, K. J., & Bradlow, A. R. (2007). Sentence recognition in native- and foreign-language multi-talker background noise. *The Journal of the Acoustical Society of America, 121*(1), 519–526.

Viswanathan, N., Kokkinakis, K., & Williams, B. T. (2016). Spatially separating language masker from target results in spatial and linguistic masking release. *The Journal of the Acoustical Society of America, 140*(6), EL465.

Weisstein, E. W. (2022). Nyquist frequency. *Https://mathworld.wolfram.com/*. https://mathworld.wolfram.com/NyquistFrequency.html

Williams, B. T., & Viswanathan, N. (2020). The effects of target-masker sex mismatch on linguistic release from masking. *The Journal of the Acoustical Society of America, 148*(4), 2006.

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.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.