**The time-course of linguistic interference during native and non-native speech-in-speech listening**

Alex Mepham1, Yifei Bi2, and Sven L. Mattys1

1 Department of Psychology, University of York

United Kingdom

2 College of Foreign Languages, University of Shanghai for Science and Technology

China

Running head: Time-course of linguistic interference

Corresponding author: Alex Mepham, Department of Psychology, University of York, York, YO10 5DD, UK. Email: am2050@york.ac.uk, ORCID: 0000-0002-6377-3719.

**ABSTRACT**

Recognizing speech in a noisy background is harder when the background is time-forward than time-reversed speech, a *masker direction effect*, and when the masker is in a known than an unknown language, indicating *linguistic interference*. We examined the masker direction effect when the masker was a known vs. unknown language, and calculated performance over 50 trials to assess differential masker adaptation. In Experiment 1, native English listeners transcribing English sentences showed a larger masker direction effect with English than Mandarin maskers. In Experiment 2, Mandarin non-native speakers of English transcribing Mandarin sentences showed a mirror pattern. Both experiments thus support the *target-masker linguistic similarity hypothesis*, where interference is maximal when target and masker languages are the same. In Experiment 3, Mandarin non-native speakers of English transcribing English sentences showed comparable results for English and Mandarin maskers. Non-native listening is therefore consistent with the *known-language interference hypothesis*, where interference is maximal when the masker language is known to the listener, whether or not it matches the target language. A trial-by-trial analysis showed that the masker direction effect increased over time during native listening but not during non-native listening. The results indicate different target-to-masker streaming strategies during native and non-native speech-in-speech listening.

Keywords: informational masking, linguistic interference, learning, non-native listening

1. **INTRODUCTION**

# Experiments investigating the ‘Cocktail Party Effect’ (Cherry, 1953) have sought to disentangle the effects of different types of maskers on the recognition of target speech. For speech-in-speech listening, it is generally agreed that challenges can arise from energetic masking, whereby a target signal is degraded due to spectro-temporal overlap with competing speech at the cochlear and auditory-nerve levels (e.g., Culling & Stone, 2017) or from informational masking, whereby target recognition is compromised by masking that is non-energetic in nature. Informational masking includes misallocations of acoustic elements from the masker to the target due to perceptual similarity, heightened cognitive load incurred by selective tracking of the target, and interference from the linguistic (e.g., phonetic, semantic) content of the masker (e.g., Cooke et al., 2008; Kidd & Colburn, 2017; Shinn-Cunningham, 2008; Summers & Roberts, 2020). Informational masking resulting from the linguistic content of the masker, which we call “linguistic interference,” is the topic of this study.

Investigations into linguistic interference have highlighted two key findings. First, a masker in a language known to the listener is often found to be more disruptive than a masker in an unknown language, which is referred to as the *known-language interference hypothesis* (Garcia-Lecumberri & Cooke, 2006; Van Engen & Bradlow, 2007; but see Mattys, Carroll, Li, & Chan, 2010, for an exception). Second, and partly overlapping with the previous hypothesis, speech recognition is usually worse if the masker language is phonetically similar to the target language, whether the masker language is known or unknown to the listener, which is referred to as the *target-masker linguistic similarity hypothesis* (Brouwer et al., 2012; Calandruccio et al., 2013; Freyman et al., 1999; Van Engen & Bradlow, 2007). For example, Brouwer et al. showed that, when listening to English sentences, both English monolingual speakers and Dutch non-native speakers of English experienced greater disruption when the competing speech was the same language as the target (English) than another language (Dutch). However, the effect was smaller for the Dutch than the English listeners, which suggests that linguistic interference is determined by an interaction between challenges with linguistic similarity between target and masker and familiarity with both the target and the masker languages.

Most studies investigating the known-language hypothesis and the language-similarity hypothesis have based their conclusions on data averaged over large numbers of trials within an experimental session. However, this overlooks the fact that the ability to stream one voice from another may change with practice and growing familiarity with the masker (Bent et al., 2009; Erb et al., 2012, 2013), and hence, average performance might misrepresent the mechanisms underlying linguistic interference. Although the time-course of speech-in-speech masking has been explored for word position within a sentence (Ezzatian et al., 2012; 2015), little is known about how performance changes across trials in the course of an experiment. One exception is Bent et al.’s (2009) study, in which the authors compared the recognition of natural speech in multi-talker babble with noise-vocoded speech over 100 sentences. They found an improvement in performance over time for both conditions. However, performance plateaued after around forty sentences in the babble condition, whereas it continued to improve up to around sixty sentences in the noise-vocoded condition. This result suggests that the ability to learn how to process degraded speech depends on the nature of the degradation (Mattys et al., 2012), with a longer and broader learning window when the degradation consists of systematic and predictable alterations of the signal (noise-vocoded speech) than when the degradation is extrinsic to the signal and mostly random (multi-talker babble). However, in the Bent et al. study, the purely linguistic properties of the masker could not be isolated because the energetic masking component of the masker was not controlled across conditions. Assessing the time-course of linguistic interference is important because it provides an insight into the ease with which a target talker can be streamed from a masker, as well as a listener’s ability to overcome the activation of the linguistic content of the masker.

Linguistic interference is usually measured as a decrease in the number of target words correctly transcribed. However, a correlate of linguistic interference (and informational masking in general) is that words from the masker are likely to be erroneously reported as belonging to the target speech through either involuntary incorporation of masker keywords into the target sentence or, in some rare cases, mistaking the masker stream for the target stream. Therefore, the incidence of masker-to-target intrusions ought to be measured alongside correct transcription performance. How these two measures develop over the course of an experiment can refine our understanding of the dynamics of linguistic interference and, specifically, listeners’ streaming improvement from trial to trial. Furthermore, comparing the ratio of target word transcription to masker word intrusion in native and non-native listeners can help pinpoint the mechanisms of disruption in these two groups when they engage in speech-in-speech perception.

In sum, the aim of this study was to improve our understanding of how linguistic interference impacts native and non-native listening over time. We ran three speech-in-speech experiments, one with native English speakers (Experiment 1) and two with Mandarin non-native speakers of English (Experiments 2 and 3). For Experiments 1 and 3, the target speech consisted of English sentences and the competing speech consisted of two-talker babble in English or Mandarin. For Experiment 2, the target speech consisted of Mandarin sentences, which were translations of the English target sentences. To minimize voice differences between the English and Mandarin babble conditions, a single English-Mandarin bilingual speaker was recorded for both conditions. Furthermore, the two-talker babble maskers in each language were created by digitally altering the fundamental frequency and vocal tract length of that speaker. Thus, all four babble voices (two English and two Mandarin) originated from a single speaker.

Finally, time-reversed versions of the English and Mandarin two-talker babble were used as a way of minimizing energetic differences between the two languages. Time-reversed speech preserves the long-term average frequency spectrum of the original signal but removes its semantic content (Licklider & Miller, 1951). Thus, the difference in performance between time-forward and time-reversed maskers provides a measure of the time-forward maskers’ ability to interfere with target recognition while controlling, to a large extent at least, for their long-term average energetic content (but see Rhebergen et al., 2005, for some limitations).

In the following experiments, we aimed to evaluate how known and unknown language maskers interfere with native vs. non-native speech recognition and how these effects develop in the course of an experiment. Specifically, we asked: (1) Is linguistic interference best accounted for by a known-language account or a linguistic similarity account? (2) Does linguistic interference change in the course of a test block, an indication of listeners’ evolving streaming capacity as familiarity with the input increases? (3) Are the above patterns affected by whether listeners perform the task in their native language as opposed to a non-native language?

# **EXPERIMENT 1: Native English listeners** **– English target sentences**

# **METHODS**

## ***Participants***

Forty native British English speakers (34 female) aged between 18 and 25 years (*M* = 20.3, *SD* = 1.9) with no known history of hearing impairments (as determined by self-report) participated in the experiment. Of those, two were excluded due to prior experience with Mandarin. The remaining 38 declared no knowledge of Mandarin or other Sinitic languages. Six of them were excluded due to technical errors during data collection. Thus, 32 participants (27 female) aged between 18 and 25 years (*M* = 20.1, *SD* = 1.7) completed the experiment and were included in the analyses. All but three of them declared knowledge of at least one other language (French, *n* = 16; Spanish, *n* = 14; German, *n* = 4; Greek, *n* = 1; Korean, *n* = 1; Welsh, *n* = 1). The University of York Department of Psychology ethics committee approved all experimental procedures for this experiment and Experiments 2 and 3 (reference number: 747). Listeners either participated for course credit or were compensated at a rate of 6.00 GBP per hour. All participants provided written-informed consent before the start of the study.

## ***Materials***

### *a. Target stimuli.* Two-hundred sentences adapted from the first 20 Harvard/IEEE sentence lists (IEEE, 1969), spoken by a female native British English speaker, were used as target stimuli (see Appendix A; all appendices can be found following the OSF link in the Acknowledgements section). Each target sentence had five keywords (e.g., “The PLAY SEEMS DULL and QUITE STUPID”, keywords capitalized). Sentence duration ranged from 1.59 s to 3.16 s (*M* = 2.20 s, *SD* = .24 s). The fundamental frequency (F0) and associated vocal tract length (VTL) of all target sentences were adjusted to a mean F0 of 210 Hz, which is approximately 15 Hz below and above the F0 of the two maskers. We manipulated the F0 and VTL of the target sentences so that the target sentences could not be distinguishable from the masker sentences solely on the basis of potential sound quality differences associated with the manipulation—by design, the masker sentences necessitated F0 and VTL alteration (see the Masker section for details). Manipulating the F0 and VTL of the target sentences also allowed us to use the same F0 and VTL values across all experiments.

### *b. Masker stimuli.* A female native Mandarin-English bilingual speaker recorded the English and Mandarin sentences used as maskers. The use of a single speaker for both sets of sentences allowed us to minimize voice variation, and hence, differences in energetic masking across conditions. Although the first language of the bilingual speaker was Mandarin, she grew up in a multilingual environment. At the time of recording, she had lived in the United Kingdom for six years.

A pilot experiment was undertaken to assess the perceived nativeness of the bilingual speaker when speaking English in relation to nine female native monolingual speakers of British English and four female Mandarin-English bilingual speakers. Twenty native English speakers were asked to judge how confident they were that each speaker grew up speaking English in the United Kingdom, using a five-point Likert scale (Not at all confident = 0, Slightly confident = .25, Somewhat confident = .50, Fairly confident = .75, Completely confident = 1.00). All listeners heard the same five sentences spoken by all speakers. Sentences were presented in a random order in a self-paced online experiment using Gorilla (Anwyl-Irvine et al., 2020). Ratings for the nine monolingual speakers, the four bilingual speakers, and the test speaker were entered in a one-way analysis of variance, which showed a significant effect of the Language Status of the Speaker (monolingual, bilingual, test speaker), *F*(2, 57) = 124.3, *p* < .001. Bonferroni-corrected post-hoc pairwise comparisons indicated that raters were more confident that the test speaker grew up speaking English in the UK (*M* = .515, *SD* = .172) than the bilingual speakers (*M* = .133, *SD* = .129, *p* < .001), a desirable feature for our experiment. However, they judged the test speaker as less likely to have grown up speaking English in the UK than the monolinguals (*M* = .842, *SD* = .127, *p* < .001).

To test the last result further, we asked a new set of twenty native English speakers to rate a single sentence (“Pack the kits and don’t forget the salt.”) produced by the test speaker and by the nine native English speakers. A single sentence was used to keep the test short. The sentence was chosen randomly. As before, participants were asked to judge how confident they were that each speaker grew up speaking English in the UK. All 10 renditions were presented as clickable icons on a computer screen, next to their corresponding Likert scale. The side-by-side format allowed the listeners to compare the renditions directly, focusing only on accentedness. Their position on the screen was randomized for each rater. Bonferroni-corrected pairwise comparisons did not show significant rating differences (*p*s > .05) between the test speaker and any but one (*p* = .009) of the native British English speakers. Thus, on balance, the results of the two tests suggest that, despite some indication that the test speaker might be detectable by some listeners as not having grown up speaking English in the UK, her speech was perceived as less accented than that of control bilinguals and as native by many listeners in our sample. We therefore judged that the test speaker’s voice was suitable to use for the masker stimuli.

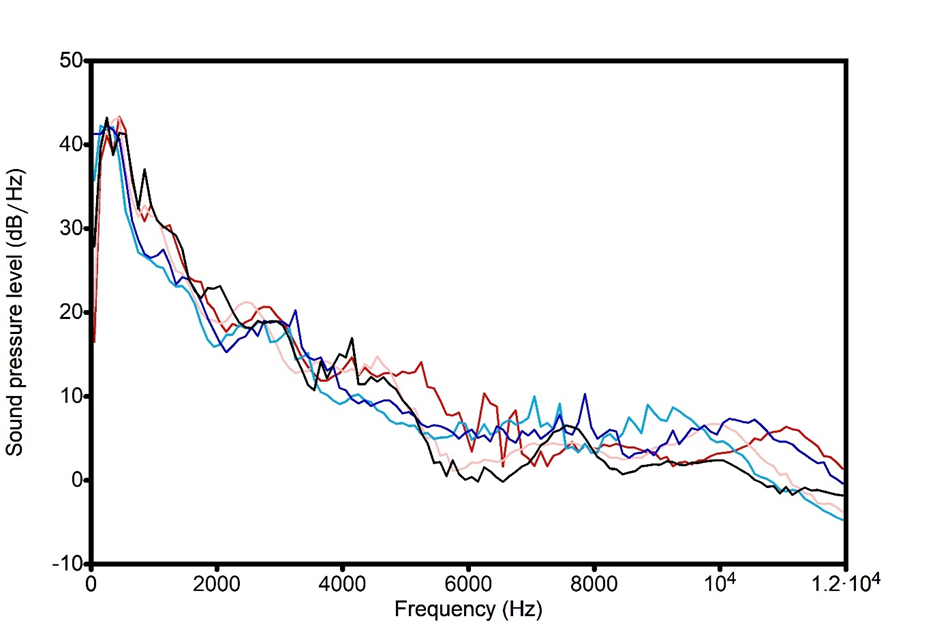
The bilingual speaker recorded 64 sentences from Lists 1-4 of the English BKB-R corpus (Bench et al., 1979). BKB-R sentences are simple sentences with three to four keywords (e.g., “The POSTMAN SHUT the GATE”, keywords capitalized). A full list of the BKB-R sentences used in this study can be found in Appendix B. The bilingual speaker also recorded Mandarin-translated versions of those sentences. Both sets of sentences were identical to those used in the Calandruccio et al. (2010) study. All sentences were recorded in a single-walled sound-attenuated room at a 44.1 kHz sampling rate with 16-bit resolution using the Audacity software. Each sentence was recorded a minimum of four times. For each sentence, the two best exemplars were kept. These constituted Set A and Set B. All sentences were manually edited using Praat (Boersma & Weenik, 2019) to remove silences at the beginning and end.

The Set A sentences were concatenated into a continuous stream, henceforth Stream A. The same was done with the Set B sentences, henceforth Stream B. Sentence order within each BKB-R list was the same in both streams, but the order of the lists differed in each stream. Both streams were edited using a version of the manipulation described in Darwin et al. (2003) to adjust the F0 and VTL (Smith et al., 2007; Gaudrain et al., 2009) so that each stream came in a high-F0 version (mean 225 Hz) and a low-F0 version (mean 195 Hz). These values were chosen to be equidistant to those of the target speaker. F0 is known to be a powerful grouping factor for speaker segregation in multi-talker environments (e.g., Bird & Darwin, 1998; Brokx & Nooteboom, 1982; Summers & Leek, 1998). Therefore, this procedure was used to control the long-term energetic overlap between the target and the maskers, i.e., how easy targets and maskers were to stream out from one another based on F0, and between the maskers themselves. As we applied the same manipulations to all three experiments in this study, the results were also more directly comparable across experiments. VTL was manipulated alongside F0 to improve the naturalness of the two streams, as both indices have been shown to contribute to the perception of voice identity (e.g., Skuk & Schweinberger, 2014). For each language, the high-F0 version of Stream A was combined with the low-F0 version of Stream B to constitute two-talker babble Version 1. Likewise, the low-F0 version of Stream A was combined with the high-F0 version of Stream B to constitute two-talker babble Version 2.

### *c.* *Long-term average spectra (LTAS).* Figure 1 displays the LTAS for the masker voices relative to the target voice. Note that, although the spectra of the target and masker voices are broadly comparable, this does not preclude the existence of local energetic masking differences between conditions. However, despite these spectral differences, the largely overlapping spectra suggest that our single-speaker procedure was effective in attenuating differences in average energetic profiles across conditions.

### *d. Target-masker mixtures.* A 500 ms silent interval was added to the onset of each target sentence. Each two-talker masker stream was segmented into 4.16 s portions, which covered the 500 ms silent onset interval, the duration of the longest target sentence (3.16 s), and an additional 500 ms silent offset interval. To create time-reversed versions of the maskers, the maskers were time-reversed from offset to onset. The time-reversed maskers allowed us to minimize potential energetic masking differences between the English and Mandarin time-forward masker conditions, and hence, provide a measure of linguistic interference that is not overly contaminated by energetic masking. Thus, this experiment included four masker conditions: (1) English time-forward, (2) English time-reversed, (3) Mandarin time-forward, and (4) Mandarin time-reversed.

### Combining target sentences and maskers was done using Praat (Boersma & Weenik, 2019). The intensity of the stimuli was normalized to 65 dB SPL for the target sentences and to 68 dB SPL for the masker streams, yielding an SNR of -3 dB, as in one of Calandruccio et al.’s (2010) experiments. The target-masker mixtures were presented diotically.



Target sentences  
English high-F0 maskers  
English low-F0 maskers  
Mandarin high-F0 maskers  
Mandarin low-F0 maskers

FIG. 1. (Color online) Long-Term Average Spectra (LTAS) averaged across target sentences, high-F0 and low-F0 English masker sentences, and high-F0 and low-F0 Mandarin masker sentences.

1. ***Procedure***

Listeners sat in a single-walled sound-attenuated booth. The experiment was conducted using PsychoPy (Pierce et al., 2019). Target-masker mixtures were presented via Denon DJ DN-HP700 headphones. Listeners were instructed to pay attention to the target speaker and to transcribe what she said using a computer keyboard. Before the main experiment started, listeners heard five practice sentences with no masker talkers to familiarize themselves with the target voice.

The main experiment had four blocks, which varied according to the masker language (English vs. Mandarin) and the masker direction (time-forward vs. time-reversed). The order of the four blocks was counterbalanced between the listeners so that, across all listeners, each condition was presented in the same block position an equal number of times, and each condition followed and preceded each other condition an equal number of times. Within each block, 50 target sentences were randomized and the random order was different for each participant. In total, each listener transcribed 200 target sentences.

Listeners transcribed the target sentences using a computer keyboard. Their transcriptions were visible on a computer monitor as they typed and they had the opportunity to delete and re-enter their responses before proceeding to the next trial. The task was self-paced. The next trial began as soon as a response was submitted.

# **RESULTS**

Listeners’ transcriptions were scored by two independent judges (authors A.M. and Y.B.) after the experiment. Obvious typographical errors were corrected. Inter-rater discrepancies (< 1% of all trials) were discussed and a score for each discrepancy was agreed upon.

Transcription scoring rules were as follows: (1) Homophones were scored as correct, e.g., “threw” and “through”, (2) Verb conjugation changes and noun pluralization were scored as correct, e.g., “have” and “had”, “cat” and “cats”, (3) Changes in syntactic category were scored as incorrect, e.g., “the parked lorry” and “the park lorry” (adjective to noun), “apart” and “part” (adverb to noun/verb), (4) Misspelt items that were real words were scored as incorrect, e.g., “rang” and “range”, (5) Misspelt items that were homophonous with the target word were scored as correct, e.g., “urge” and “urdge.” For each target sentence, listeners’ transcriptions were scored as a proportion of keywords correctly transcribed.

Participants were removed from the analysis if their mean score across all four conditions was below 0.2, indicative of generally poor performance (less than one word out of five). In Experiment 1, no participants had to be removed on that basis.

## ***Transcription performance***

Mean transcription performance by masker condition is displayed in Figure 2. The same results are broken down over time, from trial 1 to trial 50, separately for the English and Mandarin maskers in Figures 3A-B.

The data were analyzed with generalized linear mixed-effects regression models (GLMM) using the *glmer* function from the *lme4* package in R 4.0.1. All models used a binomial distribution and a logit link, with proportion of keywords correct as the dependent variable. The BOBYQA optimizer (Baayen et al., 2008) was used to aid model convergence. The base model included only the random effects, in which both listeners and stimuli were entered as random intercepts. Including slope terms to the random structure was attempted but prevented the model from converging, suggesting that these structures were over-fitted or too complex (Bates et al., 2018). The main effects of Masker Language (English, Mandarin), Masker Direction (time-forward, time-reversed), and Time (trials 1 to 50, rescaled linearly as values between 0 and 1 to assist model convergence, i.e., trial 1 corresponds to 0.02, trial 2 corresponds to 0.04, etc.) were assessed by testing the improvement in model fit when each factor, considered individually, was added to the base model. Each interaction term was assessed by comparing a model with and a model without the interaction term. Improved fit between models was estimated using likelihood ratio tests.

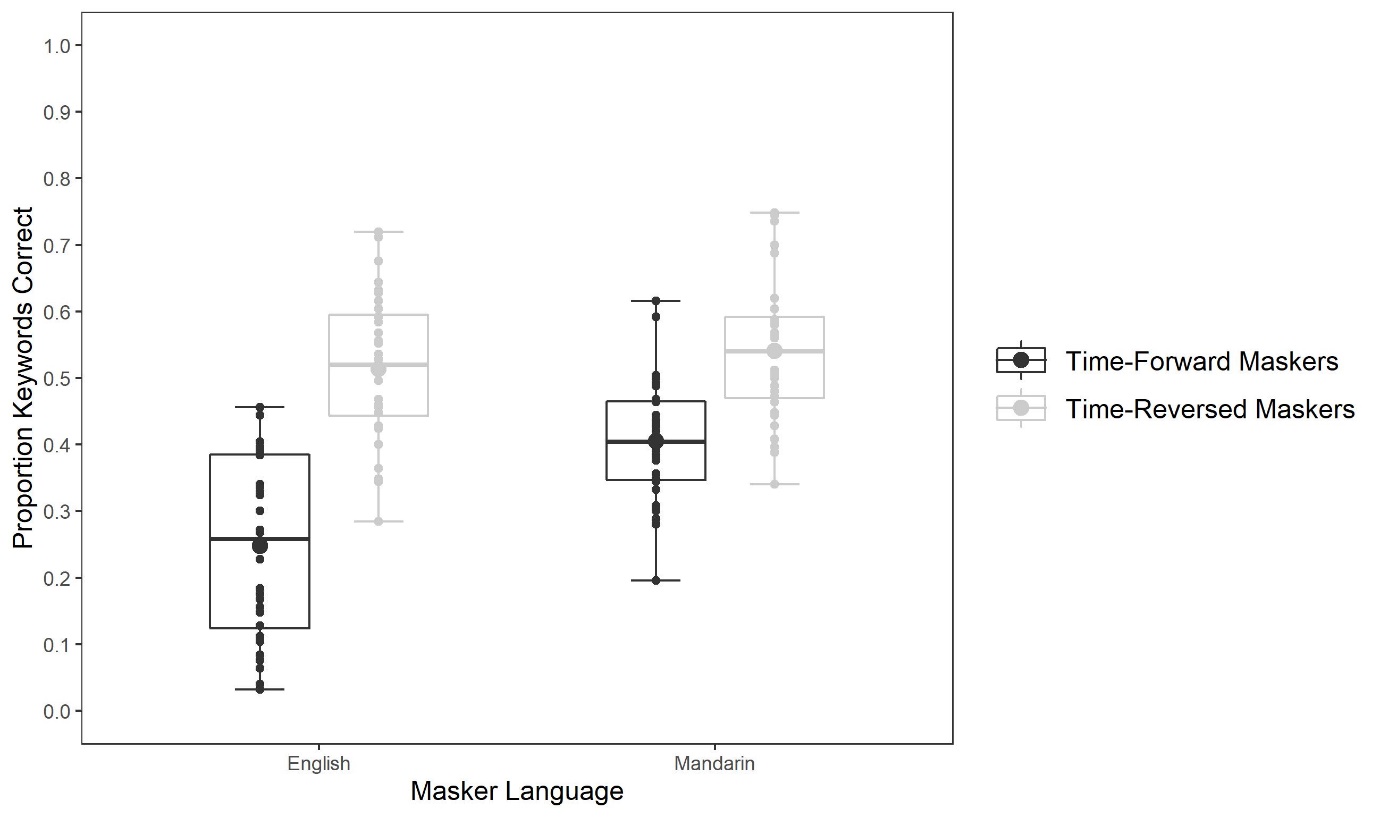


FIG. 2. Boxplot of proportion of keywords transcribed correctly as a function of Masker Language (English, Mandarin) and Masker Direction (time-forward, time-reversed). The boxes represent the interquartile range (IQR), the whiskers show 1.5 IQR over the third quartile (upper) and 1.5 IQR under the first quartile (lower), large dots represent the mean for each condition, thick horizontal bars represent the median values, and smaller dots show individual listeners.

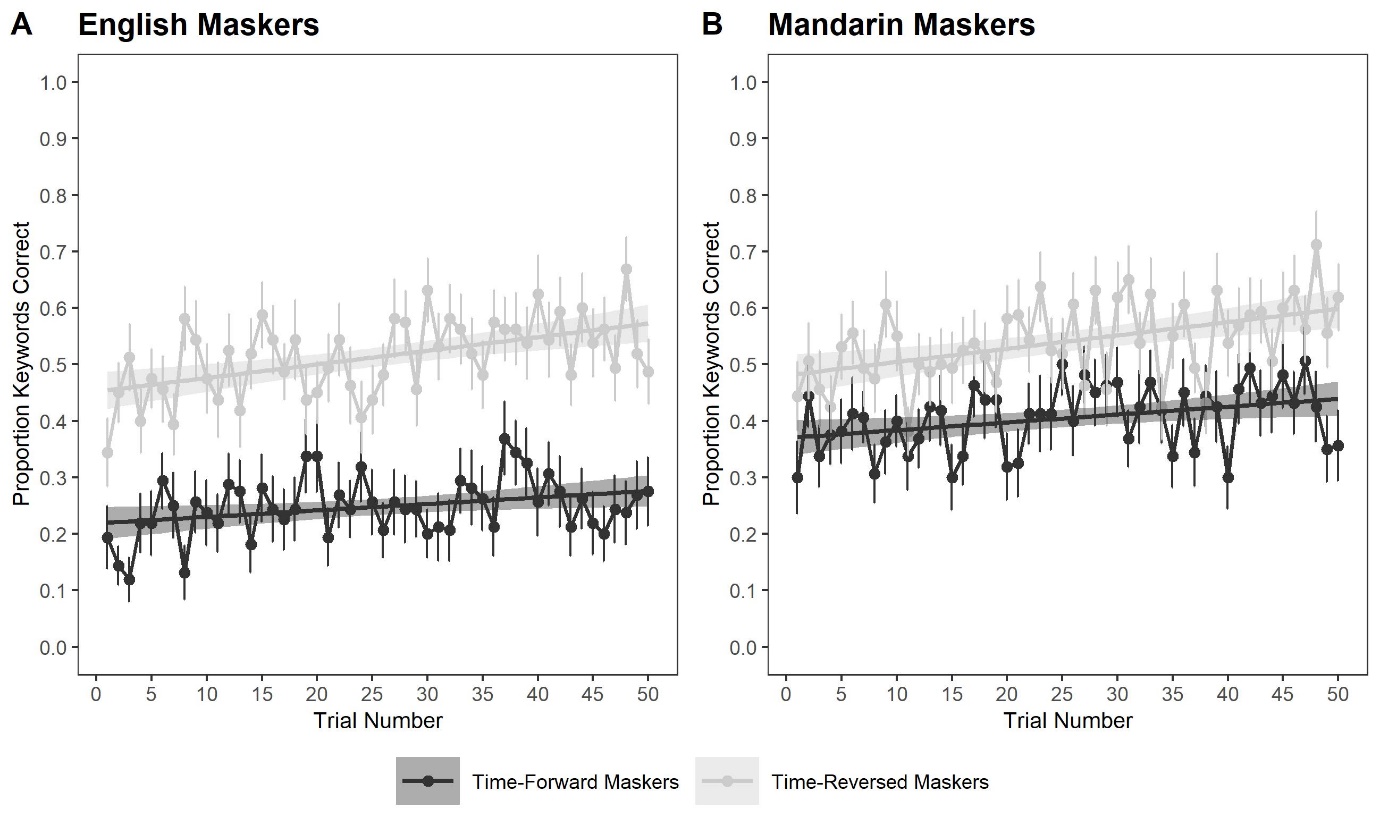


FIG. 3. Mean proportion of correct keywords for the time-forward and time-reversed masker conditions as a function of time (trials 1 to 50) for the English (A) and Mandarin (B) maskers. Error bars represent the standard error from the mean for each trial, and the shaded areas around the linear trend line represent the confidence intervals of the model fit.

Model comparison showed a Masker Language effect, with better transcription performance with Mandarin maskers (*M* = .473, *SD* = .342) than English maskers (*M* = .381, *SD* = .357), *β* = .527, *SE* = .115, *χ2*(1) = 20.68, *p* < .001, and a Masker Direction effect, with better performance for time-reversed maskers (*M* = .527, *SD* = .343) than time-forward maskers (*M* = .326, *SD* = .333), *β* = -1.06, *SE* = .104, *χ2*(1) = 93.24, *p* < .001. Performance improved over time (Time effect), *β* = .471, *SE* = .046, *χ2*(1) = 105.67, *p* < .001.

A significant interaction between Masker Language and Masker Direction, *β* = .727, *SE* = .199, *χ2*(1) = 13.17, *p* < .001, showed that the masker direction effect was larger when the masker was English than Mandarin. This is an indication of linguistic interference: Once long-term energetic differences were accounted for (i.e., the masker direction effect), the known masker (English) was more detrimental to target recognition than the unknown masker (Mandarin). Bonferroni-corrected post-hoc pairwise comparisons indicated that all four masker conditions differed from one another (all *p*s < .001), except for time-reversed English vs. time-reversed Mandarin (*p* > .05).

A significant interaction between Masker Direction and Time, *β* = -.214, *SE* = .092, *χ2*(1) = 5.41, *p* = .020, showed that, although both time-forward and time-reversed masker conditions improved over time (time-forward: *β* = .462, *SE* = .154, *χ2*(1) = 8.88, *p* = .003; time-reversed: *β* = .734, *SE* = .148, *χ2*(1) = 24.85, *p* < .001), the time-reversed condition did so more steeply. Thus, listeners found it easier to segregate the time-reversed maskers, compared to the time-forward maskers, as they progressed through the block. The three-way interaction between Masker Language, Masker Direction, and Time was not significant, *χ2*(1) = .05, *p* = .817, suggesting that this increase was comparable for both masker languages. The Masker Language by Time interaction was also non-significant, *χ2*(1) = .07, *p* = .800.

## ***Intrusion errors***

To analyze intrusions from the masker sentences into the transcription responses, we calculated the proportion of keywords originating from the masker sentences that were incorrectly reported in the listeners’ transcriptions. For example, if none of the keywords of the two masker voices were reported in the transcription, the intrusion proportion was zero. If one word out of the eight keywords in the masker voices (four in masker voice 1 + four in masker voice 2) was reported in the transcription, the intrusion proportion was .125, etc. Masker keywords were scored using the BKB-R keyword list (Bench et al., 1979) following the same rules as those used for scoring the target keywords. The intrusion analysis was restricted to the time-forward English masker condition, where the target and masker languages were the same.

Using the model-comparison approach described earlier, we assessed the effect of Time on the proportion of intrusions. Overall, the proportion of intrusions was relatively low in absolute terms (*M* = .087, SD = .160), but comparatively high when set against the correct transcription scores, which probably reflects the challenge of separating target from masker speech when the masker is in the same language as the target and intelligible to the listener. Intrusions decreased over time, *χ2*(1) = 126.68, *p* < .001. This trend is plotted in Figure 4. The correct transcription scores for that condition are plotted as well, but for reference only since the two measures are not independent from each other. The proportion of words correctly transcribed for the time-forward English masker condition did not increase significantly over time, *β* = .266, *SE* = .241, *χ2*(1) = 1.12, *p* = .290.

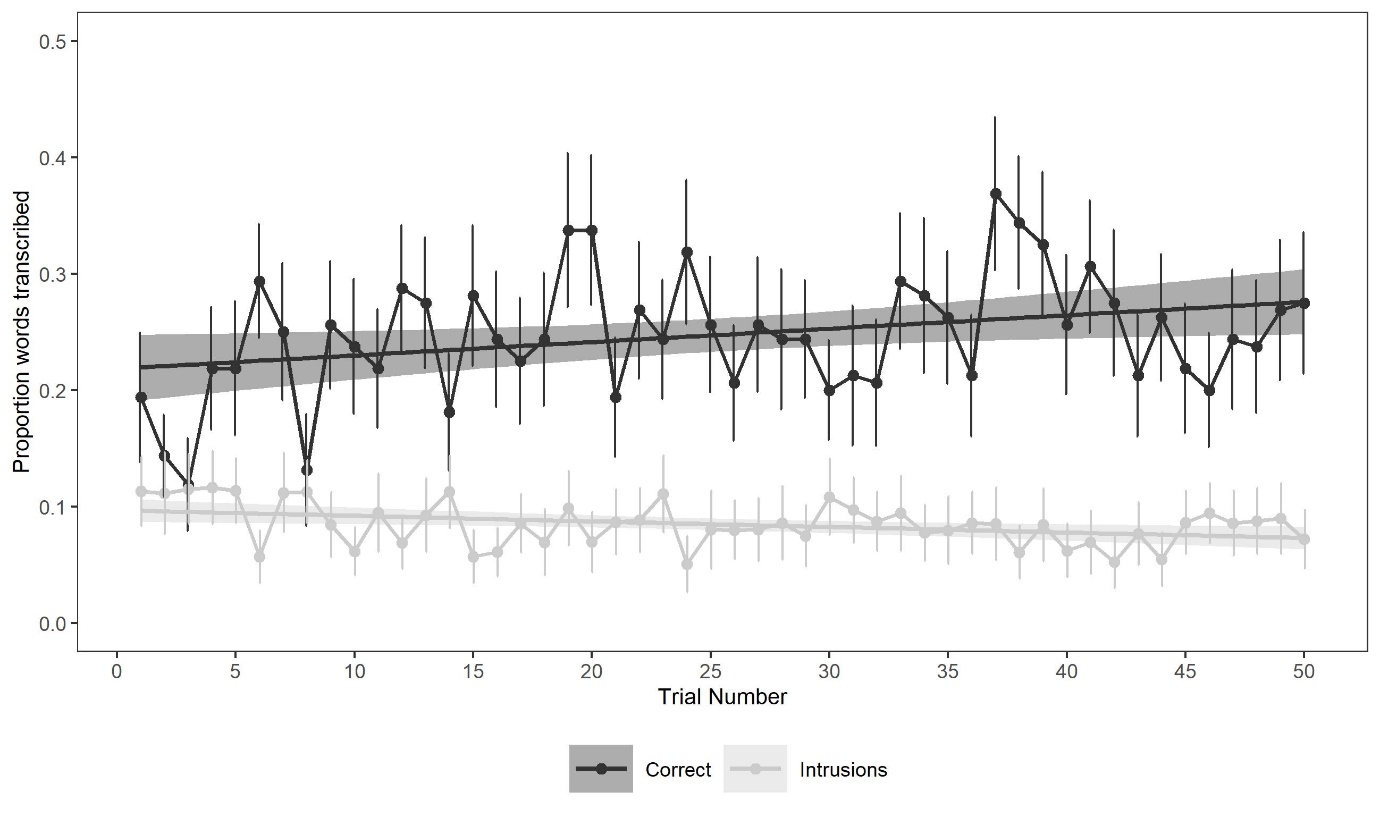


FIG. 4. Proportions of intrusions and correctly transcribed words in the time-forward English masker condition as a function of Time.

# **DISCUSSION**

This experiment shows that time-forward maskers were more disruptive than time-reversed maskers, a clear demonstration that it is harder to inhibit a masker perceived as well-formed speech (whether or not it is a known language) than a masker that violates natural speech patterns (e.g., Rhebergen et al., 2005). Critically, this masker direction effect was larger when the masker was in English than in Mandarin, which provides a strong illustration of linguistic interference and is consistent with the finding that a masker in a known language is more detrimental than a masker in an unknown language (e.g., Calandruccio et al., 2013; Cooke et al., 2008; Garcia-Lecumberri & Cooke, 2006; Van Engen & Bradlow, 2007).

Listeners were able to learn how to segregate targets from maskers over the course of the experiment, but the learning trajectory varied across conditions. Improvement was faster with time-reversed than time-forward maskers, suggesting that learning to stream and inhibit a masker was easier when the masker did not display natural speech properties (time-reversed speech) than when it did (time-forward speech). The occurrence of intrusion errors in the time-forward English masker condition is further evidence that the informational content of a masker can interfere with perception. However, the decrease in the rate of intrusions over time indicates that practice can help listeners overcome the distracting nature of the masker speech.

Although Experiment 1 provides a strong demonstration of greater interference from a known than unknown masker language, lending support to the known-language hypothesis, it does not necessarily rule out the language-similarity hypothesis. Indeed, the masker language that generated most interference (English) was both known to the listeners and similar (in fact, identical) to the target language. Conversely, the Mandarin masker was both unknown to the listeners and different from the target language.

Experiment 2 attempted to tease apart these two hypotheses. In Experiment 2, the listeners were native Mandarin speakers with non-native knowledge of English. The target sentences were Mandarin translations of the English sentences in Experiment 1. The masker stimuli were the same English and Mandarin sentences (time-forward and time-reversed) as in Experiment 1. According to the known-language hypothesis, the masker direction effect should be comparable for English and Mandarin maskers since both languages are known to the listeners—even though one is non-native. In contrast, according to the language similarity hypothesis, the masker direction effect should be greater for the Mandarin than the English maskers since the Mandarin maskers are in the same language as the target sentences, whereas the English maskers are not. A strength of this design is that, like in Experiment 1, participants performed the task in their native language. An additional strength is that the design rests on the expectation that, should the language-similarity hypothesis be correct, the results should be a mirror image of those in Experiment 1, thus decoupling the hypothesis itself from specific sets of stimuli.

# **EXPERIMENT 2: Native Mandarin listeners (with non-native English knowledge) – Mandarin target sentences**

# **METHODS**

## ***Participants***

This experiment was conducted online (see Procedure for details). Thirty-six native Mandarin speakers (29 female) aged between 18 and 35 years (*M* = 26.9, *SD* = 4.8, two missing data points), with no known history of hearing impairments, completed the experiment. Listeners’ Overall Band Score on the International English Language Testing System (IELTS) or equivalent English proficiency tests (e.g., TOEFL) was collected as a proxy measure of English proficiency. Self-reported IELTS scores ranged from 4.5 (Limited User) to 7.5 (Good User, IELTS, 2020), with a median of 6.5 (Competent User). Thirteen participants also had experience with languages other than English (Japanese, *n* = 5; French, *n* = 3; Spanish, *n* = 3; German, *n* = 2; Cantonese, *n* = 1; Catalan, *n* = 1; Shanghainese, *n* = 1). At the time of testing, 10 participants were based in the People’s Republic of China and the remaining participants were based in other countries (United Kingdom, *n* = 16; Canada, *n* = 4; Australia, *n* = 2; United States, *n* = 1; Belgium, *n* = 1; France, *n* = 1; Spain, *n* = 1; Sweden, *n* = 1). Listeners were given the choice to participate in the experiment for either a UK Amazon voucher worth 6.00 GBP or a payment of 6.00 GBP through Prolific. All participants provided written-informed consent to take part in this study.

## ***Materials***

### *a. Target stimuli.* The 200 Harvard/IEEE sentences used in Experiment 1 were translated into Mandarin and spoken by a female native-Mandarin speaker (author Y.B.). Some keywords were altered to align with everyday Mandarin. A full list of the Mandarin-translated Harvard/IEEE sentences can be found in Appendix C. Each target sentence had five keywords, as in Experiment 1. Sentence duration ranged from 1.43 s to 3.99 s (*M* = 2.55 s, *SD* = .48 s). As in Experiment 1, the F0 and VTL of all target sentences were adjusted to a mean F0 of 210 Hz.

### *b. Masker stimuli.* These were those of Experiment 1.

### *c. Target-masker mixtures.* A 500 ms silent interval was added to the onset of each target sentence. Each two-talker masker stream was segmented into 4.99 s portions, which covered the 500 ms silent onset interval, the duration of the longest target sentence (3.99 s), and an additional 500 ms silent offset interval. The maskers were time-reversed from offset to onset to create the backward masker conditions. The intensity of the target sentences was normalized to 65 dB SPL, and that of the masker streams to 68 dB SPL, yielding an SNR of -3 dB, as in Experiment 1. The target-masker mixtures were presented diotically. As in Experiment 1, Experiment 2 included four masker conditions: (1) English time-forward, (2) English time-reversed, (3) Mandarin time-forward, and (4) Mandarin time-reversed.

1. ***Procedure***

The experiment was conducted online using Gorilla (Anwyl-Irvine et al., 2020) and all instructions were presented in Mandarin. Listeners were instructed to wear headphones for the duration of the experiment. Two implementations of the experiment were created, one for listeners participating through Prolific, and one for listeners participating for UK Amazon vouchers. Listeners using Prolific were instructed to provide their Prolific ID and the listeners participating for UK Amazon vouchers were instructed to provide their email address. Listeners’ email addresses were stored separately from their data and were deleted following receipt of the Amazon voucher. Participants were asked if their first language was Mandarin or Cantonese and the experiment was terminated if they selected Cantonese.

Listeners were then requested to complete a demographic questionnaire and to calibrate their headphones to a comfortable level while listening to the first 20 s of a piece of classical music. To ensure that listeners were wearing headphones, we ran a headphone check following Woods et al.’s (2017) procedure. On each trial, three tones were presented, one of which was quieter than the others. Listeners were asked to select the quietest of the three tones. The use of antiphase audio for some of the tones meant that this task could only be successfully completed with stereo headphones. Six trials were presented, and listeners had to score at least 5/6 to continue with the study and were given two attempts to achieve this score. Following the headphone check, listeners were presented with instructions to allow autoplay of audio files in their web browser before the experiment began.

Listeners were requested to transcribe the target sentences in Chinese characters using their computer keyboard. The rest of the procedure was the same as in Experiment 1.

1. **RESULTS**

Listeners’ transcriptions were scored using a Python script (available in Appendix D). As in Experiment 1, listeners’ transcriptions were scored as a proportion of keywords correctly transcribed and all responses were checked by a native Mandarin speaker (author Y.B.). No participants were omitted as all of them achieved an average score higher than .2.

1. ***Transcription performance***

Transcription performance by masker condition is displayed in Figure 5 and broken down over time separately for the English and Mandarin maskers in Figures 6A-B.

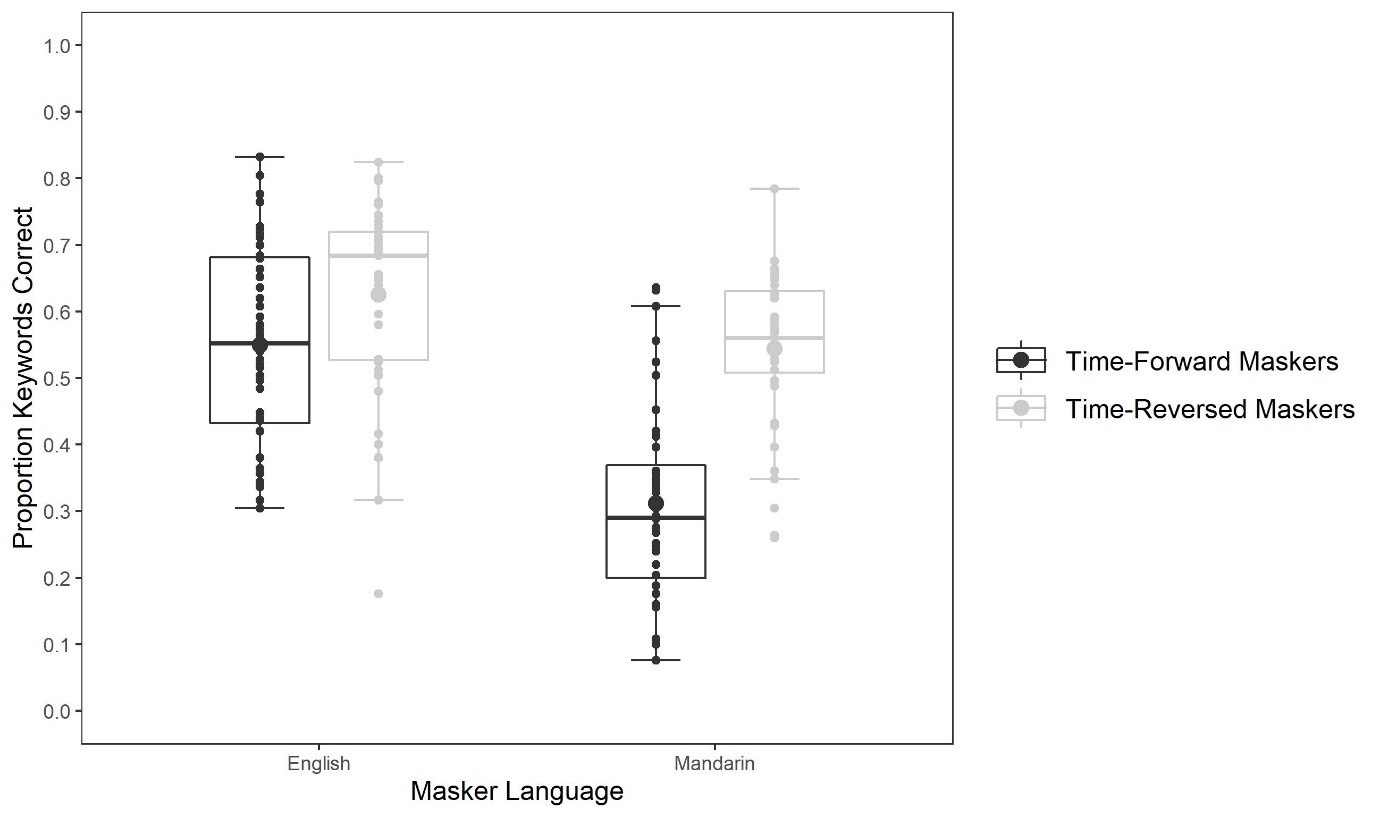


FIG. 5. Boxplot of proportion of keywords transcribed correctly as a function of Masker Language (English, Mandarin) and Masker Direction (time-forward, time-reversed).

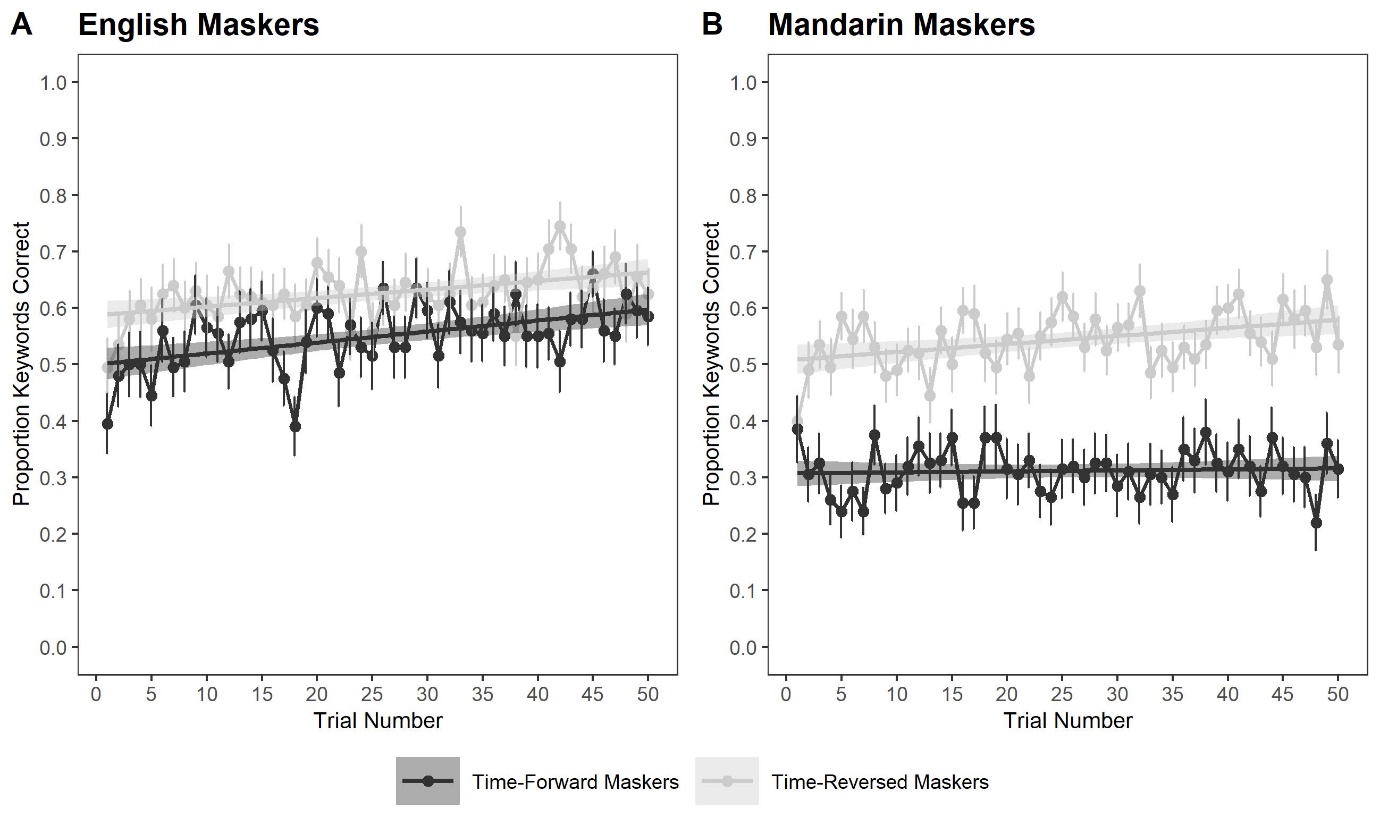


FIG. 6. Mean proportion of correct keywords for the time-forward and time-reversed masker conditions as a function of time (trials 1 to 50) for the English (A) and Mandarin (B) Masker Language conditions.

Transcription performance was analyzed following the model-comparison approach used in Experiment 1. We assessed the main effects of Masker Language (English, Mandarin), Masker Direction (time-forward, time-reversed), Time, and IELTS scores (rescaled between -0.5 and 0.5 to assist model convergence). Transcription performance was poorer with Mandarin maskers (*M* = .424, *SD* = .329) than English maskers (*M* = .578, *SD* = .331), *β* = -.580, *SE* = .080, *χ2*(1) = 49.28, *p* < .001, and poorer for time-forward maskers (*M* = .422, *SD* = .350) than time-reversed maskers (*M* = .577, *SD* = .308), *β* = .584, *SE* = .080, *χ2*(1) = 49.88, *p* < .001. Performance improved over time, *β* = .418, *SE* = .043, *χ2*(1) = 92.83, *p* < .001. There was no effect of IELTS scores, *β* = 114, *SE* = .468, *χ2*(1) = 0.06, *p* = .808.

A significant interaction between Masker Language and Masker Direction, *β* = .3946, *SE* = .103, *χ2*(1) = 14.30, *p* < .001, showed that, although the masker direction effect was significant in both masker languages (English: *β* = .479, *SE* = .130, *χ2*(1) = 13.25, *p* < .001; Mandarin: *β* = 1.130, *SE* = .133, *χ2*(1) = 64.05, *p* < .001), it was larger when the masker was Mandarin, an indication of linguistic interference, and a mirror image of the results in Experiment 1. Bonferroni-corrected post-hoc pairwise comparisons indicated that all four masker conditions differed from one another (all *p*s < .001), except for time-forward English vs. time-reversed Mandarin (*p* > .05).

A significant interaction between Masker Language and Time, *β* = -.165, *SE* = .061, *χ2*(1) = 7.22, *p* = .007, showed that performance increased faster with English than Mandarin maskers. There was no interaction between Masker Direction and Time, *β* = .008, *SE* = .061, *χ2*(1) = .02, *p* = .900. However, a significant three-way interaction between Masker Language, Masker Direction, and Time, *β* = .286, *SE* = .088, *χ2*(1) = 10.78, *p* = .001, revealed a contrast in how the masker direction effect developed over time in the English vs. Mandarin masker conditions. In the Mandarin masker condition, an interaction between Masker Direction and Time showed that the masker direction effect increased over time, *β* = .220, *SE* = .087, *χ2*(1) = 6.35, *p* = .012. This pattern was driven by an improvement in performance in the time-reversed condition, *β* = .668, *SE* = .212, *χ2*(1) = 10.07, *p* = .002, but not in the time-forward condition, *β* = .293, *SE* = .224, *χ2*(1) = 1.72, *p* = .190. In the English masker condition, an interaction between Masker Direction and Time revealed a small decrease in the masker direction effect over time, *β* = -.183, *SE* = .088, *χ2*(1) = 4.35, *p* = .037. Performance improved in both conditions (time-forward: *β* = .990, *SE* = .213, *χ2*(1) = 22.08, *p* < .001; time-reversed: *β* = .728, *SE* = .216, *χ2*(1) = 11.50, *p* <.001), but improvement was slightly faster in the time-forward condition.

***Intrusion errors***

The intrusion analysis was restricted to the time-forward Mandarin masker condition, where the target and masker languages were the same. The proportions of intrusions and correctly transcribed words are plotted in Figure 7. The proportion of intrusions (*M* = .108, *SD* = .114) was similar to that in Experiment 1, but increased over time, *β* = .155, *SE* = .070, *χ2*(1) = 4.85, *p* = .028.

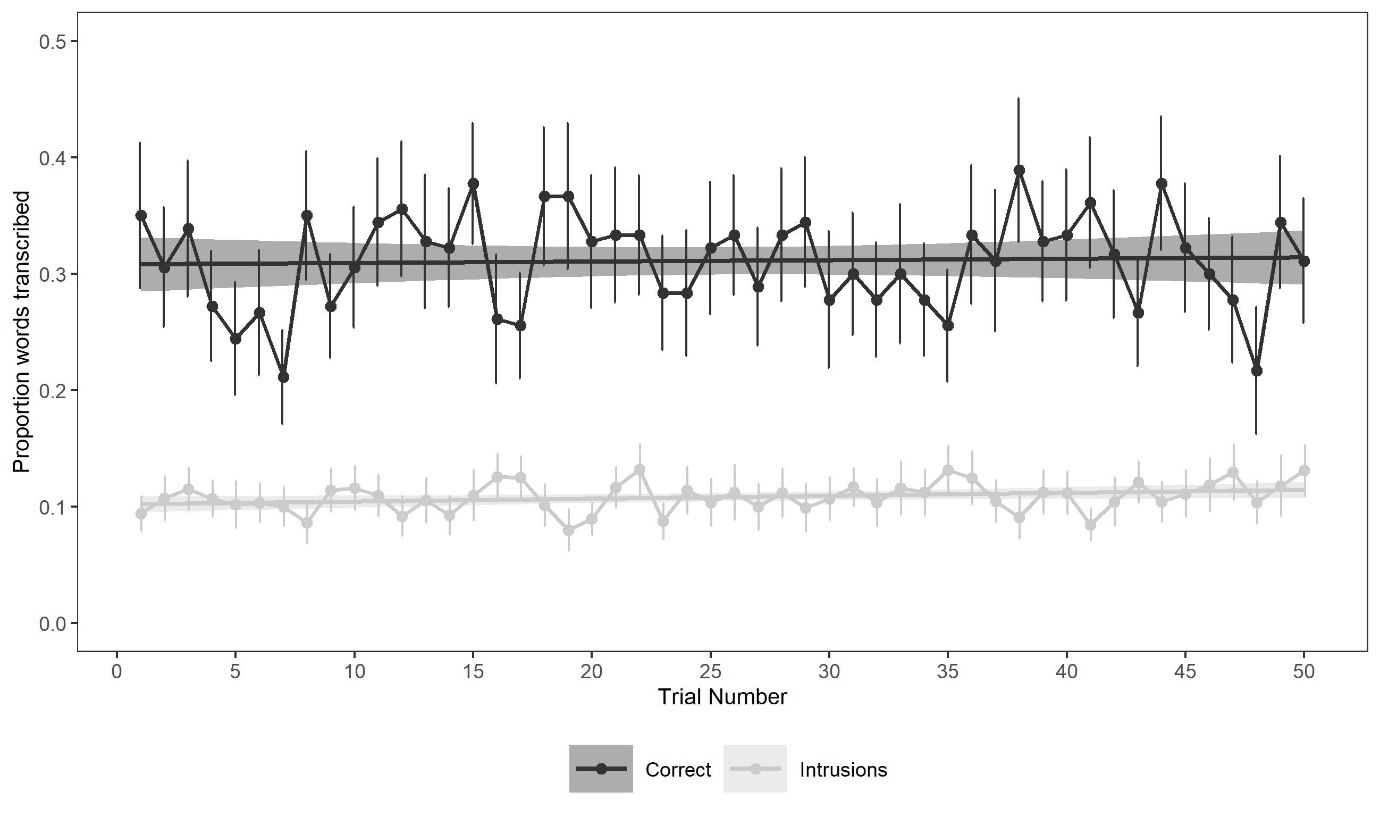


FIG. 7. Proportions of intrusions and correctly transcribed words in the time-forward Mandarin masker condition as a function of Time.

1. **DISCUSSION**

The results of Experiment 2 confirm and extend those of Experiment 1. First, the significant interaction between Masker Language and Masker Direction suggests that listeners experienced most linguistic interference when targets and maskers were in the same language (Mandarin). This result, which shows a mirror pattern to that in Experiment 1, provides a strong demonstration of linguistic interference independent of the test languages being used. Second, the results are, by and large, more compatible with the language-similarity hypothesis than with the known-language hypothesis. Linguistic interference was driven by whether the target and masker languages matched, not by whether the masker language was known to the listeners. Had performance been driven by the listeners’ knowledge of the masker language, both maskers should have led to comparable patterns of results. The distinct effect of the two masker languages was also visible in how performance changed over time: The masker direction effect increased over time for the Mandarin masker, whereas it decreased slightly for the English masker. Thus, it was harder to learn how to overcome interference from a target-matched masker (Mandarin) than from a target-mismatched masker (English), lending further support to the language-similarity hypothesis.

It could be argued that greater interference from Mandarin than English maskers occurred not because of an overlap between target and masker languages, but because listeners performed the task in their native language, and that their proficiency was better for Mandarin than for English. In Experiment 3, we explored this possibility by testing listeners with non-native knowledge of the target language. Native Mandarin speakers with non-native knowledge of English were tested on the English target sentences of Experiment 1, with the same English and Mandarin maskers as in Experiments 1 and 2. Should Experiment 3 show greater interference from English than Mandarin maskers (as in Experiment 1), this would suggest that the language similarity hypothesis holds for any language known to the listener, whether it is native or non-native. However, should a different pattern emerge, we would have to conclude that the language-similarity hypothesis is restricted to native listening and that a more complex model must be considered for non-native listening.

# **EXPERIMENT 3: Native Mandarin listeners (with non-native English knowledge) – English target sentences**

# **METHODS**

## ***Participants***

This experiment was conducted online. Thirty-two native Mandarin speakers (26 female, five male, one did not disclose) aged between 19 and 34 years (*M* = 26.1, *SD* = 4.0) with no known history of hearing impairments completed the experiment. All participants had non-native knowledge of English. All but three of them had lived in the UK—from nine months to 10 years 11 months (*n* = 21, *M* = 3.7, *SD* = 3.2). Unlike in Experiment 2, in which all participants provided an English proficiency score, not all of them did in Experiment 3. For listeners who declared their IELTS score (*n* = 21), the median was 7.0 (Good User), ranging from 6.0 (Competent User) to 8.5 (Very Good User; IELTS, 2020). IELTS scores were slightly higher in this experiment (*M* = 6.98, *SD* = 0.78) than in Experiment 2 (*M* = 6.33, *SD* = 0.95), *t*(55) = 2.76, *p* = .008. Thirteen participants declared knowledge of at least one other language in addition to English (Japanese, *n* = 6; French, *n* = 3; German, *n* = 3; Spanish, *n* = 3; Malay, *n* = 2; Cantonese, *n* = 1; Portuguese, *n* =1, Russian, *n* = 1). Listeners either participated in this experiment for course credit or were compensated at a rate of 6.00 GBP per hour. All participants provided written-informed consent to take part in the study.

## ***Materials***

### The target and masker stimuli were those of Experiment 1. However, to account for performance differences previously reported between native and non-native listening (Bradlow & Alexander, 2007; Brouwer et al., 2012), the intensity of the masker sentences was lowered to a more favorable SNR. After piloting various SNRs, the intensity of the target speech was set to 59 dB SPL. The target sentences were played at 65 dB SPL, as in Experiments 1 and 2. Thus, in Experiment 3, the SNR was +6 dB, compared to -3 dB in Experiments 1 and 2. The target-masker mixtures were presented diotically.

## ***Procedure***

The procedure was identical to that of Experiment 2.

# **RESULTS**

## ***Transcription performance***

Transcription performance by masker condition is displayed in Figure 8 and broken down over time separately for the English and Mandarin maskers in Figures 9A-B. Two participants with mean performance lower than .2 were removed from subsequent analyses (*n* = 30).

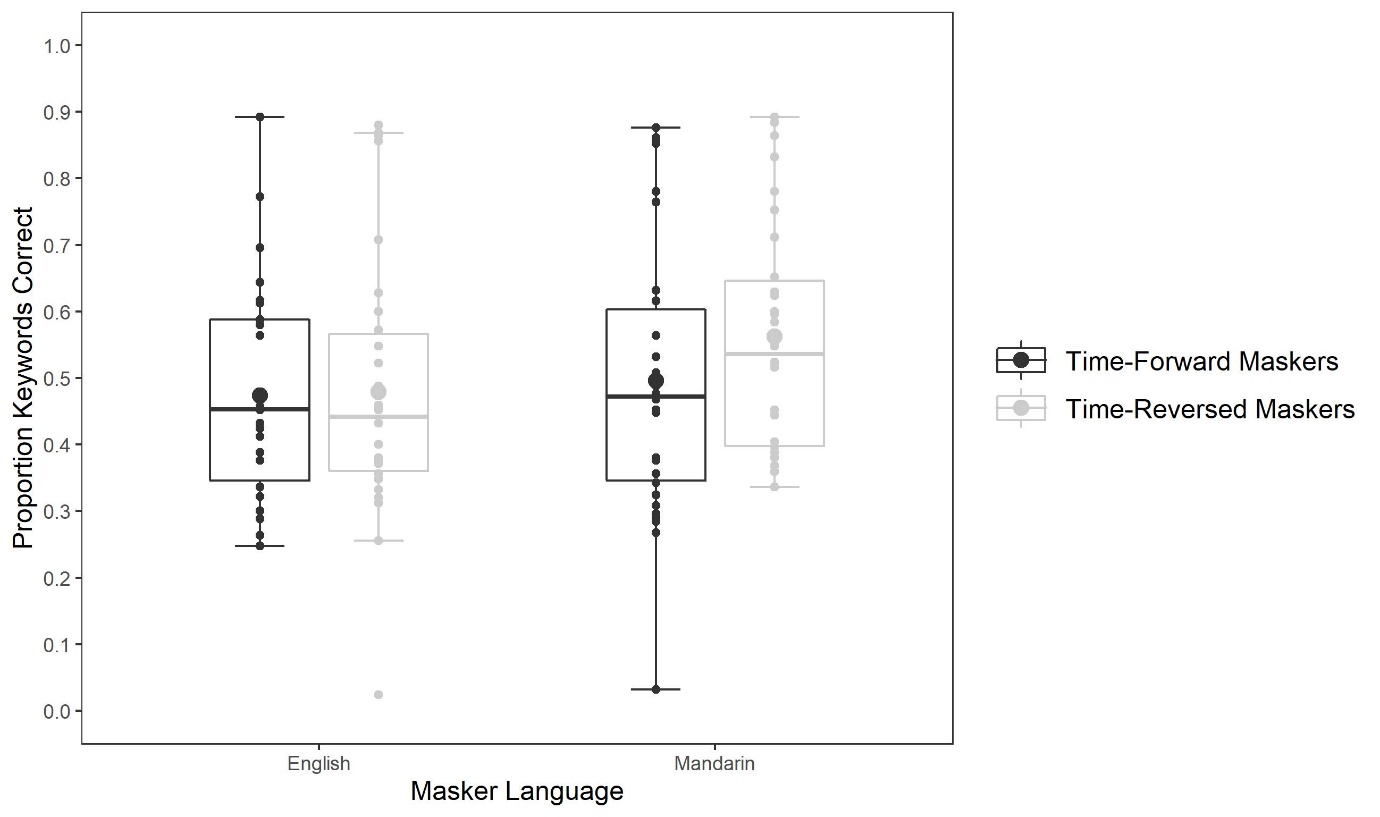


FIG. 8. Boxplot of proportion of keywords transcribed correctly as a function of Masker Language (English, Mandarin) and Masker Direction (time-forward, time-reversed).

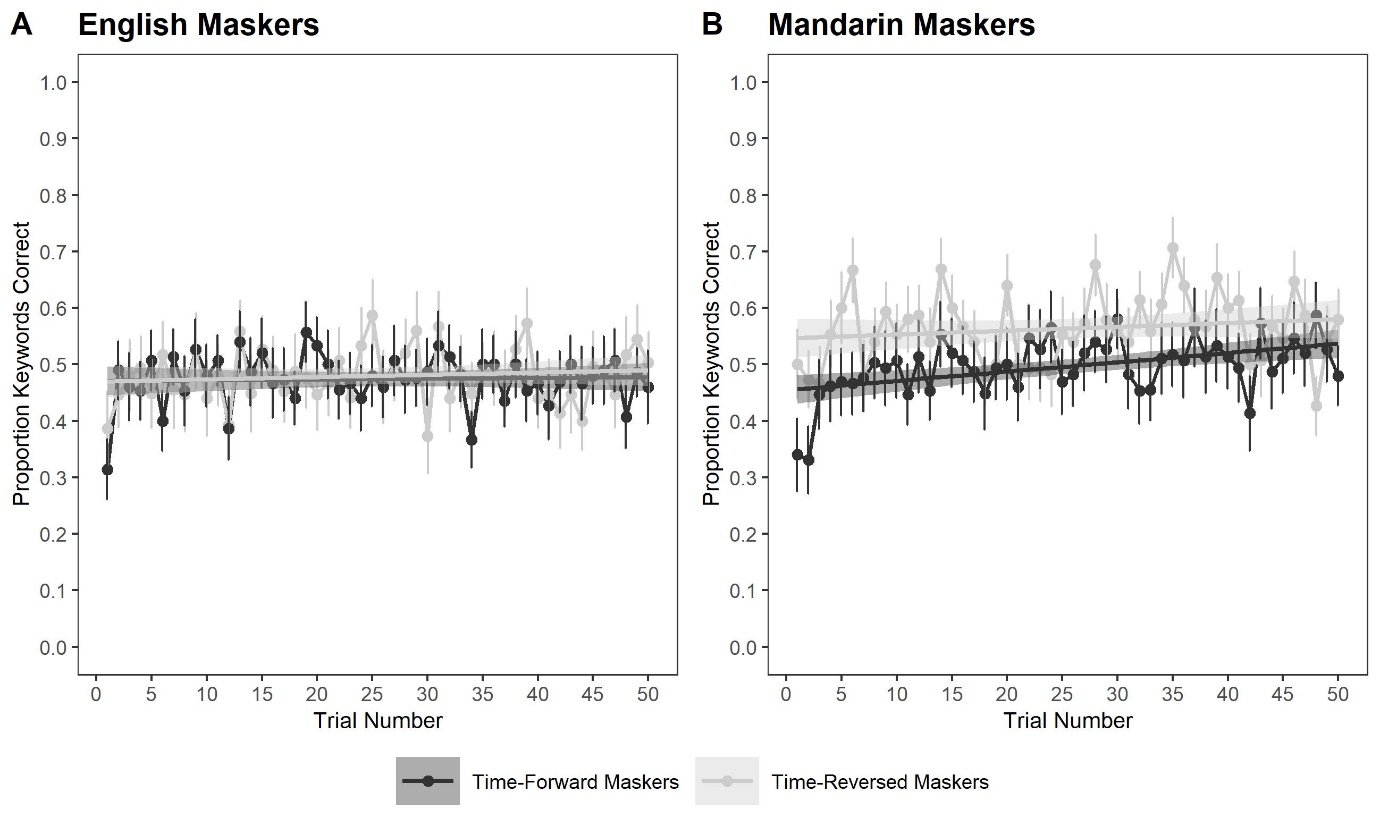


FIG. 9. Mean proportion of correct keywords for the time-forward and time-reversed masker conditions as a function of time (trials 1 to 50) for the English (A) and Mandarin (B) Masker Language conditions.

We assessed the main effects of Masker Language (English, Mandarin), Masker Direction (time-forward, time-reversed), and Time, with listeners and stimuli as random intercepts. Transcription performance was poorer with Mandarin (*M* = .53, *SD* = .19) than English maskers (*M* = .48, *SD* = .18), *β* = .300, *SE* = .091, *χ2*(1) = 10.81, *p* = .001, and poorer with time-forward (*M* = .48, *SD* = .19) than time-reversed maskers (*M* = .52, *SD* = .19), *β* = -.221, *SE* = .091, *χ2*(1) = 5.80, *p* = .016. Performance improved over time, *β* = .147, *SE* = .047, *χ2*(1) = 9.59, *p* = .002. None of the two-way interactions reached significance (all *p*s > .05), but there was a significant three-way interaction, *β* = .416, *SE* = .189, *χ2*(1) = 4.83, *p* = .028: The Masker Direction by Time interaction was significant in the Mandarin masker condition, *β* = .349, *SE* = .135, *χ2*(1) = 6.66, *p* = .010, but not in the English masker condition, *β* = -.081, *SE* = .133, *χ2*(1) = 0.37, *p* = .545. Thus, the masker direction effect decreased over time in the Mandarin masker condition, with performance improving in the time-forward condition, *β* = .413, *SE* = .096, *χ2*(1) = 18.36, *p* < .001, but not in the time-reversed condition, *β* = .048, *SE* = .096, *χ2*(1) = 0.25, *p* = .617. In the English masker condition, neither the time-forward, *β* = .036, *SE* = .094, *χ2*(1) = 0.15, *p* = .700, nor the time-reversed condition, *β* = .122, *SE* = .097, *χ2*(1) = 1.55, *p* = .213, changed with time.

In an attempt to account for English proficiency, the data were re-analyzed for the listeners who provided IELTS scores (21 minus the two participants with performance scores < .2, *n* =19). Only results involving the IELTS factor are reported. An effect of IELTS scores showed that high IELTS scores were associated with better transcription performance, *β* = 1.59, *SE* = .401, *χ2*(1) = 11.46, *p* < .001. However, an interaction between IELTS scores and Masker Language, *β* = .0.312, *SE* = .112, *χ2*(1) = 7.73, *p* = .005, showed that this association was only present in the English masker condition. There was also a significant four-way interaction between Masker Language, Masker Direction, Time, and IELTS scores, *β* = -1.914, *SE* = .814, *χ2*(1) = 5.46, *p* = .019. However, the patterns arising from this interaction did not lend themselves to a straightforward interpretation, probably due to the small sample size, and are therefore not reported.

## ***Intrusion errors***

It was not possible to model the intrusion data in the English Forward condition due to their extremely low occurrence (less than 2%).

# **DISCUSSION**

Unlike Experiments 1 and 2, Experiment 3 did not show a marked contrast between English and Mandarin maskers in terms of the masker direction effect. In other words, for the non-native listeners of Experiment 3, there was no evidence of linguistic interference. Instead, for that group, the results are more consistent with the idea that the masker direction effect is determined by whether the masker language is known to the listener, as per the known-language hypothesis (Garcia-Lecumberri & Cooke, 2006; Van Engen & Bradlow, 2007) rather than by whether the masker language matches the target language, as per the language-similarity hypothesis (Brouwer et al., 2012; Calandruccio et al., 2013; Freyman et al., 1999; Van Engen & Bradlow, 2007). Had the latter been true, the masker direction effect would have been greater in the English than Mandarin masker condition, as was the case for the native English listeners in Experiment 1 and (in a mirror fashion) for the native Mandarin listeners in Experiment 2. It could be argued that the wide individual differences in Experiment 3 compared to Experiments 1 and 2 made the relevant interaction more difficult to find. However, a visual inspection of the data (and the three-way interaction) in Experiment 3 shows that the language direction effect was, in fact, numerically larger for the Mandarin masker condition, rather than for the English masker condition. Therefore, if anything, there is evidence that the masker more likely to interfere with the task was the masker known natively by the listener rather the masker overlapping with the target language.

Another way in which the non-native listeners in Experiment 3 differed from the native listeners in Experiments 1 and 2 was the relatively minor effect that time had on performance. While performance generally improved over the course of Experiment 3, as it did in Experiments 1 and 2, this effect was driven mainly by an improvement in the time-forward Mandarin masker condition, that is, the masker direction effect decreased as listeners adapted to the native masker. The lack of masker direction effect in the English masker condition suggests that the non-native listeners managed to inhibit the non-native masker from the very beginning of the experiment, even though the masker was the same language as the target. The very low rate of intrusions from the English masker supports that conclusion.

# **GENERAL DISCUSSION**

The aim of this study was to investigate how listeners’ knowledge of the linguistic content of a masker impedes target speech perception, and how this effect is amenable to change with practice over the course of an experiment. We tested two listener groups (native listeners in Experiments 1-2 and non-native listeners in Experiment 3) who differed in their knowledge of the linguistic content of the maskers (English vs. Mandarin maskers). English and Mandarin maskers were played in their original format (time-forward) or backward (time-reversed), with the time-reversed condition providing a baseline for any long-term spectral and energetic masking differences between the two masker languages. Linguistic interference was measured as the difference in masker direction effect (time-forward vs. time-reversed) between the English and Mandarin masker conditions.

## **Transcription Performance**

For the native listeners, the masker direction effect was largest when the masker was the same language as the target speech and when it was known to the listener (Experiment 1). This result is consistent with both the known-language hypothesis (Garcia-Lecumberri & Cooke, 2006; Van Engen & Bradlow, 2007) and the language-similarity hypothesis (Van Engen & Bradlow, 2007; Brouwer et al., 2012; Calandruccio et al., 2013, 2017). However, this pattern persisted even when the listeners had knowledge of both masker languages (Experiment 2), which suggests that the results cannot be accounted for entirely by the idea that knowledge of the masker language is the driving force behind linguistic interference, as per the known-language hypothesis. Experiment 2 rather fits with the literature showing that speech-in-speech recognition is most impaired when a masker shares speech characteristics (e.g., phonology, prosody) with the target language and, by implication, when the target and masker languages are the same, as per the language similarity hypothesis.

Our finding shows both dissimilarities and similarities with Calandruccio et al.’s (2010) study. When Calandruccio et al. used the same SNR as we did (-3 dB), they did not show greater interference from an English than Mandarin masker during native English listening. However, at a more challenging SNR (-5 dB), their results and ours aligned in showing greater interference from the English masker. Interestingly, our average performance at -3 dB SNR was within the range of their average performance at -5 dB SNR, in fact even lower. Thus, their study and ours converge in showing linguistic interference when the listening conditions are challenging. However, the two studies differ in how energetic differences between the maskers were handled. In our study, we attempted to minimize potential energetic differences by using time-reversed speech as a relative baseline, whereas Calandruccio et al. compared English and Mandarin maskers with each other directly. However, in subsequent analyses, they considered their data in the context of masker-modulated noise analogues and LTAS profiles. Those analyses confirmed that, in the easy SNR condition, energetic masking differences could account for linguistic interference, whereas, at the more challenging SNR, they could not.

In contrast with the native results of Experiments 1-2, the participants who performed the task non-natively (Experiment 3) did not show greater interference from the target-matched masker (English) than the target-mismatched masker (Mandarin). In fact, there was some evidence of greater interference in the target-mismatched condition. This suggests that, for non-native listeners, it is the (native) knowledge of the linguistic content of the masker that drives interference, rather than linguistic similarities between target and masker languages. This finding challenges the language similarity hypothesis and, instead, supports the known-language hypothesis. These data provide an interesting counterpoint to a study by Calandruccio and Zhou (2014), who assessed the recognition of English sentences against English and Greek maskers by monolingual English speakers vs. English-Greek bilinguals. They found greater interference from the English masker in both groups, which led them to conclude that it is the similarity between target and masker languages, rather than the listener’s knowledge of the masker language, that drives linguistic interference. However, a difference between their experiment and ours was that, while our participants in Experiment 3 were clearly non-native speakers of English, and could therefore be said to perform the task non-natively, the participants in Calandruccio and Zhou were simultaneous bilinguals from Greek descents who were born in the USA and started learning both English and Greek from birth or shortly after. Those participants effectively performed the task natively. The finding of a language-similarity effect in that group is therefore consistent with our claim that native listening conforms to the language-similarity hypothesis, whereas non-native listening conforms to the known-language hypothesis.

Related to this point, it is important to note that, in Experiment 3, listeners’ English proficiency appeared to have influenced transcription accuracy, with high-proficiency listeners performing better than low-proficiency listeners. Although our interpretation of the contribution of language proficiency to linguistic interference is limited by the relatively small number of IELTS scores available in our experiments, these preliminary patterns confirm the need to consider the language proficiency of listeners performing tasks in their non-native language (see Scharenborg & van Os, 2019, for a review; see von Hapsburg & Peña, 2002, for methodological considerations).

It is worth considering an alternative explanation for the smaller masker direction effect and linguistic interference during non-native than native listening. Recall that, in order to match average performance between the two conditions, we had to set the SNR at -3 dB for native listening (Experiments 1-2) and +6 dB for non-native listening (Experiment 3). It has been shown that, for native speakers at least, increased SNR is generally associated with both better performance and lower informational masking, probably because misallocation of masker content is less likely to occur when streaming is made easier by a more advantageous SNR (e.g., Arbogast et al., 2005; Freyman et al., 2008). However, if the higher SNR in the non-native listening condition had made streaming easier, one would also have expected performance to be higher. However, performance was comparable in both groups. Therefore, although SNR differences should be considered in future research, we do not think that the reduced linguistic interference and lower intrusion rate during non-native listening can be entirely explained by the higher SNR.

## **Improvement over time**

One of the goals of this study was to explore the changes in the masker direction effect and in linguistic interference over the course of the experiment. While transcription performance improved for listeners performing the task either natively (Experiments 1 and 2) or non-natively (Experiment 3), the change in the masker direction effect differed between the two groups. The native listeners showed an increase in the masker direction effect over time in most conditions, with a faster rate of improvement for time-reversed than time-forward speech. Thus, they were better at learning to suppress a masker that did not conform to natural speech patterns than one that did. However, note that, for the bilingual listeners in Experiment 2, this pattern was not found when the masker was in English, i.e., the target-mismatched masker. For that condition, improvement was slightly better for time-forward than time-reversed maskers. For listeners performing the task in a native language, experience with bilingualism might therefore confer some ability to inhibit the interference of a non-native language.

The general increase in masker direction effect for the native listeners is incompatible with the hypothesis that life-long familiarity with time-forward speech (even in an unknown or non-native language) accelerates object formation in the course of an experiment, hence sharpening *object-based auditory attention* (Shinn-Cunningham, 2008) more for time-forward than time-reversed maskers. Instead, the slower improvement with time-forward than time-reversed maskers is consistent with Bent et al.’s (2009) observation that perceptual adaptation is poorer in multi-talker babble than against noise-vocoded maskers. They ascribed this difference to the novelty of the noise-vocoded speech, and hence, its learnability over time, compared to the lower potential for learnability of an already familiar signal such as babble noise. Applied to our results, the logic would be that time-reversed speech, which is unfamiliar to most listeners, would have more learnability potential than time-forward speech. For instance, the unfamiliar spectral-temporal characteristics of certain phonemes when played backward (Rhebergen et al., 2005), while distracting at first, could be learned over time and, ultimately, make it easier to for the time-reversed speech to be interpreted as an auditory object distinct from the target speech.

Additionally, although both time-forward and time-reversed speech mixtures contained local spectro-temporal modulations that can result in opportunities to ‘glimpse’ the target speech (Brungart et al., 2006; Cooke, 2006; Festen & Plomp, 1990), local glimpsing opportunities may be different in time-forward and time-reversed maskers (Buss et al., 2020). Opportunities to glimpse the target speech in the masker stream may be enhanced in the time-reversed masker condition as there is no additional interference from the linguistic content of the masker. Regardless of whether the familiarity or the glimpsing explanation is correct, our results are at odds with the hypothesis that listeners can learn to overcome the masker direction effect with practice. Thus, the distracting nature of a well-formed masker (whether it is a known or unknown language) might be largely automatic, at least during native listening.

In contrast to the listeners performing the task in their native language (Experiments 1 and 2), the non-native listeners (Experiment 3) showed no change in the masker direction effect for the English masker and a small decrease for the native but target-mismatched masker (Mandarin). We hypothesize that the better ability to overcome the masker direction effect over time during non-native listening is the unintended consequence of the high level of effort involved in listening to a non-native language. Previous studies have shown that non-native speakers experience greater listening effort than native speakers when listening to speech in both quiet and background noise (Borghini & Hazan, 2018) and when listening to speech spoken by either native or non-native speakers (Song & Iverson, 2018). Increased effort due to non-native listening could have resulted in the listeners doing the transcription task at the limit of their cognitive capacity, exhausting the resources that native listeners might otherwise use to involuntarily process distractor information (cf. Lavie al.’s [2004] Perceptual load Theory). In such conditions, whether the masker was time-forward or time-reversed, or whether or not it matched the target language, would have had little impact. Thus, native vs. non-native differences in prioritization of cognitive resources to the main transcription task could explain why there was a large and increasing interference in Experiments 1 and 2, and minimal linguistic interference in Experiment 3—but see also considerations about SNR differences in the previous section.

1. **Intrusions**

The intrusion of masker words into listeners’ responses was analyzed in the masker condition where the target and masker languages were the same (time-forward English in Experiments 1 and 3 and time-forward Mandarin in Experiment 2). For the native English listeners (Experiment 1), intrusions decreased throughout the block while accuracy correspondingly increased. For the native Mandarin listeners (Experiment 2), intrusions showed a small increase across trials. The significant occurrence of intrusions in these two groups and the fact that they traded off or mirrored the accuracy data suggests that native listeners were unable to fully inhibit the linguistic content of the masker, even though there was some evidence that they learned to partly overcome its effects through practice. In contrast, non-native listening (Experiment 3) led to almost no intrusion errors at all, consistent with the absence of linguistic interference in the accuracy data.

Lavie et al.’s (2004) perceptual load theory can, again, be drawn upon to explain the difference in intrusions between the two groups. In this framework, high perceptual load reduces distractor interference because it exhausts the resources needed to process the relevant stimuli, leaving little capacity for processing distractor stimuli. This hypothesis has received some support from Francis’ (2010) demonstration that increasing perceptual load (from an easy to a hard tone-perception task) reduced the interference of a competing voice on target speech perception. Under the assumption that the phonology and prosody of a non-native target language constitute a situation of high perceptual load, such perceptual load would guard against interference (intrusions) from the irrelevant stimulus (masker). This could explain why the masker direction effect was small and intrusions were almost non-existent for the listeners performing the task in a non-native language compared to a native language. Heightened investment of processing resources towards the transcription task in the non-native listeners (Borghini & Hazan, 2018; Song & Iverson, 2018) and performance being possibly at its peak for that group would also mean that little spare capacity was left for improvement in transcription performance over time.

## **CONCLUSION**

This study investigated the effect of masker direction (time-forward vs. time-reversed speech) and masker language over the course of an experiment with native English speakers and native Mandarin speakers with non-native knowledge of English. Better performance with time-reversed maskers than time-forward maskers was pervasive for English and Mandarin listeners performing the task in their native language, and this masker direction effect was particularly pronounced when the masker language was the same as the target language. This result supports the target-masker linguistic similarity hypothesis (Brouwer et al., 2012; Calandruccio et al., 2013; Freyman et al., 1999; Van Engen & Bradlow, 2007), whereby speech-in-speech interference is maximal when the target and masker languages share characteristics (e.g., phonology, prosody) and, by implication, when they are the same language, as was the case here. Furthermore, for listeners performing the task in their native language, the masker direction effect increased over the course of the experiment, which suggests that it is easier to learn to inhibit time-reversed speech than time-forward speech. For listeners performing the task in a non-native language, the masker direction effect was broadly equivalent across the two known masker languages, thus supporting the known-language interference hypothesis (Garcia-Lecumberri & Cooke, 2006; Van Engen & Bradlow, 2007). There were also more intrusion errors during native than non-native listening. We hypothesize that listening in a non-native language might force listeners to engage a large proportion of their cognitive capacity toward the target speech and, as a consequence, reduce opportunities for distraction by the masker.

# **ACKNOWLEDGEMENTS**

# This research was supported by a research grant from the Leverhulme Trust (RPG-2019-152) to S.L.M. We are grateful to Sarah Knight and Lyndon L. Rakusen for their advice concerning the set-up of the experimental paradigm, and to Sarah Knight, Ronan McGarrigle, and Sophie Meekings for feedback on earlier drafts. We also thank Lauren Calandruccio and Ann R. Bradlow for providing the English BKB-R sentences and Mandarin audio files for the BKB-R sentence translations, and Sarah Knight for supplying relevant Praat scripts. We also thank Miaomiao Yu for recording the English and Mandarin BKB-R sentences and for translating the BKB-R sentences alongside Lydia Y. Stuhrmann. We are also grateful to the speakers who provided speech samples for the initial pilot study. All audio stimuli, analysis scripts, data files, and Appendices can be found at https://osf.io/nhcrw/.

**REFERENCES**

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behav. Res. Methods, 52,* 388-407. https://doi.org/10.3758/s13428-019-01237-x

Arbogast, T. L., Mason, C. R., & Kidd Jr, G. (2005). The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.*, *117*, 2169-2180. https://doi.org/10.1121/1.1861598

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modelling with crossed random effects for subjects and items. *J. Mem. Lang., 59,* 390-412. https://doi.org/10.1016/j.jml.2007.12.005

Bates, D., Kliegl, R., Vasishth, S., Baayen, R. H. (2018). Parsimonious Mixed Models. arXiv.org/abs/1506.04967

Bench, J., Kowal, Å., & Bamford, J. (1979). The Bkb (Bamford-Kowal-Bench) Sentence Lists for Partially-Hearing Children. *Br. J. Audiol., 13,* 108-112. https://doi.org/10.3109/03005367909078884

Bent, T., Buchwald, A., & Pisoni, D. B. (2009). Perceptual adaptation and intelligibility of multiple talkers for two types of degraded speech. *J. Acoust. Soc. Am., 126,* 2660-2669. https://doi.org/10.1121/1.3212930

Bird, J., & Darwin, C. J. (1998). Effects of a difference in fundamental frequency in separating two sentences. In A. R. Palmer, A. Rees, A. Q. Summerfield, & R. Meddis (Eds.), *Psychophysical and Physiological Advances in Hearing*, pp. 263-269. Whurr, London.

Boersma, P. & Weenink, D. (2019). Praat: doing phonetics by computer [Computer program]. Version 6.1, retrieved 28 September 2019 from http://www.praat.org/

Borghini, G., & Hazan, V. (2018). Listening effort during sentence processing is increased for non-native listeners: A pupillometry study. *Front. Neurosci., 12,* 1-13. https://doi.org/10.3389/fnins.2018.00152

Bradlow, A. R., & Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *J. Acoust. Soc. Am., 121,* 2339-2349. https://doi.org/10.1121/1.2642103

Brokx, J. P., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *J. Phon., 10,* 23-36. https://doi.org/10.1016/S0095-4470(19)30909-X

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 control. *J. Acoust. Soc. Am., 131,* 1449-1464. https://doi.org/10.1121/1.3675943

Brungart, D. S., Chang, P. S., Simpson, B. D., & Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *J. Acoust. Soc. Am., 120,* 4007-4018. https://doi.org/10.1121/1.2363929

Buss, E., Calandruccio, L., Oleson, J., & Leibold, L. J. (2020). Contribution of Stimulus Variability to Word Recognition in Noise Versus Two-Talker Speech for School-Age Children and Adults. *Ear & Hearing, 42,* 313-322. https://doi.org/10.1097/AUD.0000000000000951

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. *Am. J. Audiol., 22,* 157-164. https://doi.org/10.1044/1059-0889(2013/12-0072)

Calandruccio, L., Dhar, S., & Bradlow, A. R. (2010). Speech-on-speech masking with variable access to the linguistic content of the masker speech. *J. Acoust. Soc. Am., 128(2),* 860-869. https://doi.org/10.1121/1.3458857

Calandruccio, L. & Zhou, H. (2014). Increase in Speech Recognition due to Linguistic Mismatch Between Target and Masker Speech: Monolingual and Simultaneous Bilingual Performance. *J. Speech Lang. Hear. Res., 57,* 1089-1097. https://doi.org/10.1044/2013\_JSLHR-H-12-0378

Cherry, C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *J. Acoust. Soc. Am., 25,* 975-979. https://doi.org/10.1121/1.1907229

Cooke. M. (2006). A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am., 119,* 1562-1573. https://doi.org/10.1121/1.2166600

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. *J. Acoust. Soc. Am., 123,* 414-427. https://doi.org/10.1121/1.2804952

Culling, J. F., & Stone, M. A. (2017). Energetic Masking and Masking Release. In J. C. Middlebrooks, J. Z. Simon, A. N. Popper, & R. R. Fay (Eds.) *The Auditory System at the Cocktail Party* (pp. 41-73). Springer Nature. https://doi.org/10.1007/978-3-319-51662-2

Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *J. Acoust. Soc. Am., 114*, 2913-2922.

Erb, J., Henry, M. J., Eisner, F., & Obleser, J. (2012). Auditory skills and brain morphology predict individual differences in adaptation to degraded speech. *Neuropsychologica, 50,* 2154-2164. https://doi.org/10.1016/j.neuropsychologica.2012.05.013

Erb, J., Henry, M. J., Eisner, F., & Obleser, J. (2013). The brain dynamics of rapid perceptual adaptation to adverse listening conditions. *J. Neurosci., 33,* 10688-10697. https://doi.org/10.1523/JNEUROSCI.4596-12.2013

Ezzatian, P., Li, L., Pichora-Fuller, M. K., & Schneider, B. A. (2012). The effect of energetic and informational masking on the time-course of stream segregation: Evidence that streaming depends on vocal fine structure cues. *Lang. Cognitive Proc., 27,* 1056-1088. https://doi.org/10.1080/01690965.2011.591934

Ezzatian, P., Li, L., Pichora-Fuller, K., & Schneider, B. A. (2015). Delayed stream segregation in older adults: More than just informational masking. *Ear & Hearing, 36,* 482-484. https://doi.org/10.1097/AUD.0000000000000139

Festen, J. M. & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *J. Acoust. Soc. Am., 88,* 1725-1736. https://doi.org/10.1121/1.400247

Francis, A. L. (2010). Improved segregation of simultaneous talkers differentially affects perceptual and cognitive capacity demands for recognizing speech in competing speech. *Atten. Percept. Psycho.*, *72*, 501-516. https://doi.org/10.3758/APP.72.2.501

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2008). Spatial release from masking with noise-vocoded speech. *J. Acoust. Soc. Am.*, *124*, 1627-1637. https://doi.org/10.1121/1.2951964

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceptive spatial separation in the unmasking of speech. *J. Acoust. Soc. Am., 106,* 3578-3588. https://doi.org/10.1121/1.428211

Garcia Lecumberri, M. L., & Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. *J. Acoust. Soc. Am., 119,* 2445-2454. https://doi.org/10.1121/1.2180210

Gaudrain, E., Li, S., Ban, V. S., Patterson, R. (2009). The role of glottal pulse rate and vocal tract length in the perception of speaker identity. *Proceedings of* *Interspeech 2009*, Brighton, United Kingdom. https://hal.archives-ouvertes.fr/hal-02144510/document

Händel, G. F. (ca. 1731-35). Bourrée. From *Water Music HWV 348*. Dieren: Canzona Music. https://imslp.org/wiki/Special:ReverseLookup/527960, Retrieved 18 August 2020

IEEE Subcommittee on Subjective Measurements (1969). IEEE Recommended Practices for Speech Quality Measurements. *IEEE Trans. Audio Electroacoust., 17,* 227-246. https://doi.org/10.1109/TAU.1969.1162058

International English Language Testing System (IELTS; 2020). How IELTS is scored. https://www.ielts.org/about-the-test/how-ielts-is-scored

Jin, S.-H., & Liu, C. (2012). English sentence recognition in speech-shaped noise and multi-talker babble for English-, Chinese-, and Korean-native listeners. *J. Acoust. Soc. Am., 132,* EL391-EL397. https://doi.org/10.1121/1.4757730

Kidd, G. Jr., & Colburn, H. S. (2017). Informational Masking in Speech Recognition. In J. C. Middlebrooks, J. Z. Simon, A. N. Popper, & R. R. Fay (Eds.) *The Auditory System at the Cocktail Party* (pp. 75-109). Springer Nature. https://doi.org/10.1007/978-3-319-51662-2

Lavie, N., Hirst, A., de Fockert, J. W., & Viding, E. (2004). Load Theory of Selective Attention and Cognitive Control. *J. Exp. Psychol. Gen., 133,* 339-354. https://doi.org/10.1037/0096-3445.133.3.339

Licklider, J. C. R., & Miller G. A. (1951). The perception of speech. In S. S. Stevens (Ed.) *Handbook of experimental psychology* (p. 1040-1074). Wiley.

Mattys, S. L., Carroll, L. M., Li, C. K. W., & Chan, S. L. Y. (2010). Effects of energetic and informational masking on speech segmentation by native and non-native speakers. *Speech Commun., 52,* 887-899. https://doi.org/10.1016/j.specom.2010.01.005

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Lang. Cognitive Proc., 27*, 953-978. https://doi.org/10.1080/01690965.2012.705006

Pierce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51,* 195-203. https://doi.org/10.3758/s13428-018-01193-y

Rhebergen, K. S., Versfeld, N. J., & Dreschler, W. A. (2005). Release from informational masking by time-reversal of native and non-native interfering speech. *J. Acoust. Soc. Am., 118,* 1274-1277. https://doi.org/10.1121/1.2000751

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 Commun., 108,* 53-64. https://doi.org/10.1016/j.specom.2019.03.001

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends Cogn. Sci., 12,* 182-186. https://doi.org/10.1016/j.tics.2008.02.003

Skuk, V. G., & Schweinberger, S. R. (2014). Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender. *J. Speech Lang. Hear. Res.*, *57*, 285-296. https://doi.org/10.1044/1092-4388(2013/12-0314)

Smith, D. R. R., Walters, T. C., & Patterson, R. D. (2007). Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled. *J. Acoust. Soc. Am., 122,* 3628-3639. https://doi.org/10.1121/1.2799507

Song, J., & Iverson, P. (2018). Listening effort during speech perception enhances auditory lexical processing for non-native listeners and accents. *Cognition, 179,* 163-170. https://doi.org/10.1016/j.cognition.2018.06.001

Summers, V., & Leek, M. R. (1998). F0 processing and the seperation of competing speech signals by listeners with normal hearing and with hearing loss. *Journal of Speech, Language, and Hearing Research*, *41*, 1294-1306. https://doi.org/10.1044/jslhr.4106.1294

Summers, R. J., & Roberts, B. (2020). Informational masking of speech by acoustically similar intelligible and unintelligible interferers. *J. Acoust. Soc. Am.*, *147*, 1113-1125. https://doi.org/10.1121/10.0000688

Van Engen, K. J. (2010). Similarity and familiarity: Second language sentence recognition in first- and second-language multi-talker babble. *Speech Commun., 52,* 943-953. https://doi.org/10.1016/j.specom.2010.05.002

Van Engen, K. J., & Bradlow, A. R. (2007). Sentence recognition in native- and foreign-language multi-talker background noise. *J. Acoust. Soc. Am., 121*, 519-526. https://doi.org/10.1121/1.2400666

von Hapsburg, D., & Peña, E. D. (2002). Understanding bilingualism and its impact on speech audiometry. *Journal of Speech, Language, and Hearing Research, 45, 202-213.* https://doi.org/10.1044/1092-4388

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,* 2064-2072. https://10.3758/s13414-017-1361-2