Navigating the bilingual cocktail party: Interference from background speakers in listeners with varying L1/L2 proficiency

Emilia Lew, Sophie Hallot, Krista Byers-Heinlein, and Mickael Deroche

A Thesis

in the Department

of Psychology

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Arts (Psychology) at

Concordia University

Montréal, Québec, Canada

March 2023

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Emilia Lew**

Entitled: **Navigating the bilingual cocktail party: Interference from background speakers in listeners with varying L1/L2 proficiency**

and submitted in partial fulfillment of the requirements for the degree of

Master of Arts (Psychology)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

_____Chair

Dr. Andrew Chapman

_____Examiner

Dr. Emily Coffey

_____Examiner

Dr. Virginia Penhune

_____Supervisor

Dr. Mickael Deroche

Approved by

_____

Dr. Andrew Ryder, Chair of Department

_____

Dr. Pascale Sicotte, Dean, Faculty of Arts and Science

Date: March 14th, 2023

ABSTRACT

Navigating the bilingual cocktail party: Interference from background speakers in listeners with varying L1/L2 proficiency

Emilia Lew, B.A.

Cocktail party environments require listeners to tune in to a target voice while ignoring surrounding speakers (maskers), which could present unique challenges for bilingual listeners. Our study recruited English-French bilinguals to listen to a male target speaking French or English, masked by two female voices speaking French, English, or Tamil, or by speech-shaped noise. Listeners performed better with first language (L1) than second language (L2) targets, and relative L1/L2 proficiency acted like a categorical rather than a continuous variable with respect to speech reception threshold (SRT) *averaged over maskers*. Further, listeners struggled the most with L1 maskers and struggled the least with Tamil maskers. The results suggest that the balanced bilinguals have a slight disadvantage with L1 targets but compensate with a larger advantage with L2 targets, compared to unbalanced bilinguals. This positive net result supports the idea that being a balanced bilingual is helpful in speech-on-speech perception tasks in environments that offer substantial exposure to L2.

Key words: cocktail party problem, bilingualism, speech perception, competing speech task

Acknowledgements

I would like to thank my research supervisor, Dr. Mickael Deroche, for his endless patience and guidance during my time in this program. I would also like to thank the members of my committee, Dr. Emily Coffey and Dr. Virginia Penhune, as well as co-author Dr. Krista Byers-Heinlein, for their invaluable feedback throughout the writing process. Thank you as well to Sophie Hallot for getting this project started and continuing to assist us in research even after starting medical school at McGill. I am incredibly grateful to the members of the Hearing and Cognition Lab for helping me find my way around Concordia University, both literally and figuratively, and for offering me advice in times of confusion and stress. To my partner, family, and friends, thank you for your love, kindness, and endless support. I could not do this without you.

Thank you to all the participants on Prolific who gave their time to complete this study. I would also like to acknowledge the support of the Centre for Research on Brain, Language, and Music's (CRBLM) pilot grant awarded to M.D. and K.B.H. (Research Incubator Award, reference FRQ-NT RS-203287). The CRBLM is funded by the Government of Quebec via the Fonds de Recherche Nature et Technologies and Société et Culture. Finally, thank you to Ms. Ramiya Veluppillai for her help in making the trilingual recordings of the maskers used in this study.

Contribution of Authors

Mickael Deroche and Krista Byers-Heinlein conceptualized the study. Mickael Deroche and Sophie Hallot coded the online experiment. Sophie Hallot collected the data via Prolific. Emilia Lew and Sophie Hallot cleaned and organized the online data. E.L. and Mickael Deroche analyzed the data. Emilia Lew wrote the final manuscript that was then edited by Krista Byers-Heinlein and Mickael Deroche. All authors reviewed the final manuscript and approved the contents.

Table of Contents

List of Tables

Page

List of Figures

Page

1. Introduction

Imagine you are attending a crowded cocktail party, trying to hear what a friend is saying over the noise in the room. To make matters even more challenging, you happen to be at an event with international attendees speaking a variety of different languages. In this situation, would you have more trouble understanding your friend if they spoke your first language or your second? Would it matter what language the other guests were speaking? Though previous research has been conducted on speech comprehension in cocktail party situations, most of this research has restricted itself to monolingual listeners (Allen, 1994; Qian, Weng, Chang, Wang & Yu, 2018). While this approach can be useful in reducing variability between participants to examine some of the underlying mechanisms of speech comprehension and source segregation, these studies do not address the experiences of the estimated half of the global population who speak two or more languages (Grosjean, 2012). Research that has included bilingual participants had generally compared them to monolinguals (e.g., Broersma & Scharenborg, 2010; Calandruccio & Zhou, 2014) without exploring how differences amongst bilinguals relate to their performance, for example in their language proficiency (Luk, 2015; de Bruin, 2019; DeLuca, Rothman, Bialystok & Pliatsikas, 2019). Indeed, exploring such individual differences can provide insight into the processes that underly speech recognition (Bregman, 1990; Kidd, Mason, Richards, Gallun & Durlach, 2008; Bronkhorst, 2015). The current study examined English-French bilinguals' language comprehension abilities in noisy environments, varying both the target language of the speaker (L1 vs. L2) and the language of the masking voices (L1, L2, or completely foreign).

Monolinguals and competing speech tasks

Listeners' performance in cocktail party situations relates to how different auditory signals mask each other, and these phenomena generally fall under energetic and informational masking effects. Energetic masking occurs when the target and masking stimuli share similar auditory qualities, for example, when two people are talking at the same time (temporal similarity), or when the person speaking has a similar pitch range to others chatting around them (frequency similarity) (Kidd et al., 2008). Informational masking occurs when the content of distractor stimuli contributes to difficulties in comprehension of the target stimulus above and beyond what can be explained by energetic masking. This is best demonstrated with attentional tasks devoid of energetic masking (e.g., random-frequency multitone bursts, Oxenham. Fligor, Mason & Kidd, 2003) or when target and masking sentences are processed so that they occupy different spectral channels (Kidd, Mason & Gallun, 2005). Part of what makes the cocktail party problem so interesting and difficult is that it combines energetic and informational masking in the same listening environment (Brungart, Simpson, Ericson & Scott, 2001). As such, a given auditory cue such as a difference in the speakers' voice pitch or a difference in their spatial position may provide both a release from energetic and from informational masking (Hawley, Litovsky & Culling, 2004; Deroche, Culling, Lavandier & Gracco, 2017a). Differences in the speakers' language are another cue that can lead to both forms of masking release but have been relatively little investigated.

Some studies used a competing speech paradigm while varying the language of the masker and found that monolingual participants are better able to understand the (native) target voice when masking voices are speaking a foreign language, as opposed to the participants' native language (Van Engen & Bradlow, 2007; Calandruccio, Dhar & Bradlow, 2010). However,

in this situation, it is impossible to say whether native language stimuli are more challenging because of the participants' proficiency in their native language or because the masking language is the same as the target. The former interpretation points exclusively to informational masking: native sounds would elicit speech units in the brain that are more likely to resonate within the linguistic and semantic knowledge of the listener, causing greater distraction when searching for native target words. The latter interpretation suggests that greater similarity between target and masker would increase informational masking but also possibly energetic masking. For example, a native vowel is more likely than a non-native vowel to occupy spectral regions where target cues are present, and a native syllable is more likely than a non-native syllable to possess the sort of temporal modulations that are common in the rhythm of the native language. Quantifying the difference between two languages would require taking repeated audio recordings of each overlapping vowel and consonant sound (for example, the French /a/ and the English /a/), averaging them in some way (e.g. long-term spectrum or some temporal characteristics), comparing the similarity in spectro-temporal modulation across languages for these overlapping sound pairs, and then accounting for sounds that are distinct to each language (for example, the French /u/ and the English /th/). Though complicated, calculating auditory language differences is theoretically possible. Quantifying the degree of informational masking, by contrast, is infinitely more challenging, if not impossible. It is notoriously difficult to identify the nature of masking in speech-on-speech situations (Brungart, 2001) – nor is it the goal of this study – but testing the former versus the latter interpretation of this native language masking phenomenon would help and is something that is made possible by turning to bilingual listeners.

Bilinguals and competing speech tasks

Several studies have investigated bilinguals' performance in competing speech tasks, usually in comparison to monolingual participants when the target is presented in a single language (Van Engen, 2010; Calandruccio & Zhou, 2014; Regalado, Kong, Buss & Calandruccio, 2019). Van Engen (2010) and Calandruccio and Zhou (2014) found that simultaneous bilingual participants performed better when the masker language differed from the target language, consistent with the pattern observed for monolingual participants. On the other hand, Regalado et al. (2019) found that monolingual, simultaneous bilingual, and late bilingual (L1 Mandarin, L2 English) participants performed better with noise maskers as opposed to speech maskers (all targets and speech maskers were presented in English) but they did not investigate the role of masker language.

However, to compare monolingual and bilingual samples, the target voice must use the monolingual participants' L1. Most studies comparing monolingual and bilingual performance in competing speech tasks have been conducted in the US, where monolinguals are typically native English speakers and bilinguals are typically immigrants or children of immigrants. Unless the bilinguals recruited are simultaneous bilinguals, English is usually the L2 of the bilingual participants, as exemplified in Regalado et al. (2019), in which English was the L2 of late bilingual participants. Differing levels of proficiency relative to the target language both reduces the validity of the comparison between bilinguals and monolinguals and makes it difficult to determine how bilinguals might perform on a competing-speech task using their L1 as a target. Regalado et al. (2019) also noted worse performance in the late bilingual group under both masking conditions, but they determined that this was due to differences in language dominance

rather than bilingual disadvantage, further illustrating the inevitable differences in performance that arise when L1 differs between groups.

There are several unique factors that could affect bilinguals' performance in competing speech tasks. First, with respect to the target language, most bilinguals have uneven proficiency in their L1 and L2. Bilinguals with higher L2 proficiency are more likely to perform better in speech perception tasks with L2 targets (Kilman, Zekveld, Hällgren & Rönnberg, 2014; Warzybok, Brand, Wagener & Kollmeier, 2015), but exactly how much better as a function of their relative proficiency in each language is unclear. Second, with respect to the masker language, given that monolinguals perform better when the masking voice is foreign rather than native (Van Engen & Bradlow, 2006; Calandruccio et al., 2010), bilinguals might also perform better with maskers speaking in a foreign language or in a poorly mastered L2. According to Green's (1998) inhibitory control model, however, bilinguals suppress one of their languages to produce the other. According to this model, all words known to an individual contain a language "tag" linking them to a specific language. When people are speaking in a target language, the non-target language words that are simultaneously activated during the speech production process are inhibited because they have the non-target language tag. Now, one may question the neurophysiological root of such tags in the speech network of the brain, but many studies support this model, not only indicating that bilinguals engage in inhibition of one language's words while producing speech the other, but also showing that greater inhibition of one language leads to better production of the other (Linck, Hoshino & Kroll, 2008; Linck, Kroll, & Sunderman, 2009; Pivneva, Palmer & Titone, 2012; Declerck, Thoma, Koch & Philipp, 2015). These studies suggest that bilinguals might perform better in competing speech tasks where the masking language and target language differ (although they do not say whether performance would be

better in a L1 target vs L2 masker situation as opposed to a L2 target vs L1 masker situation). Inhibitory mechanisms against the non-target language may work in tandem with the participants' active intention to ignore the masking speakers to provide greater masker release than in scenarios where the target and masking language are the same. This reasoning leads to a rather surprising prediction. Instead of L1 maskers always being the most effective, as in the native masking phenomenon, the inhibitory control model suggests that performance would be better in a L2 versus L1 situation than in a L2 versus L2 situation or a L1 versus L1 situation. Furthermore, proficiency in L2 would help with L2 targets but might make L2 maskers more effective as well. It is not clear which one is more beneficial: better target intelligibility or less masking. The native masking phenomenon and the inhibitory control model lead to different predictions, partly because effects related to the intrinsic intelligibility of the target voice and effects related to masking have not been properly disentangled. As a result, it is currently unknown how these two factors would influence bilingual listening abilities in such speech-on-speech situations.

Hypotheses

Our study examines bilingual performance with L1 and L2 targets, comparing performance across L1, L2 and foreign language (Lf) maskers, thus expanding our understanding of bilingual listening behavior in cocktail party situations. With respect to target language, we predicted an effect of proficiency, such that participants would perform better with L1 targets than L2 targets, and we were curious to know whether the size of this difference would vary linearly or non-linearly as a function of their relative proficiency in the two languages. With respect to masker language, we predicted that participants would have poorer performance when the target and masker languages were the same than when they were different, but we suspected

that this pattern might not necessarily hold the same way for balanced and unbalanced bilinguals.

2. Methods

1. Participants

A total of 200 French-English bilingual participants were recruited through the Prolific platform. All participants spoke either English or French as their L1, and the other language as their L2, and were between 18 and 50 years of age. Participants were asked to report the languages they spoke, in addition to age of acquisition (AOA), listening proficiency, listening use, speaking proficiency, and speaking use for all languages. Proficiency and use metrics were self-scored from 1 to 10, with 10 indicating highest proficiency/use.

A total of 72 participants were excluded. Participants were excluded if they did not complete the study due to technical difficulties (6 participants), if they did not follow the instructions (overscoring or underscoring; 4 participants), or if their performance indicated that they did not have a high enough proficiency in L2 to understand L2 targets when these targets were perfectly audible (62 participants). This resulted in 57 participants in the L1 ENGLISH group (40 women and 17 men) and 71 participants in the L1 FRENCH group (34 women, 36 men, 1 not reported).

2. Stimuli

The English target stimuli were sourced from the Institute of Electrical and Electronics Engineers' recommended practice for speech quality measurements, often termed the Harvard sentences (IEEE, 1969). This corpus of 720 phonemically-balanced, standardized English sentences was originally created to test audio quality in various telephone systems, but has since expanded in use in psychoacoustic research. The speaker of these target stimuli was a North

American male. An example sentence from this stimuli set is, “The kite dipped and swayed, but stayed aloft.” The French target stimuli were a translation of the Harvard sentences termed the Fharvard corpus (Aubanel, Bayard, Strauß & Schwartz, 2020; openly available), also produced by a male (but French native) speaker. The Fharvard sentences were phonemically-balanced to minimize the differences in phoneme distribution between sentences. An example sentence from this stimuli set is, “La jeunesse passe toujours tellement vite.” The Harvard and FHarvard sentences were chosen because they have been validated for use in language comprehension studies. In both corpora, the stimuli were trimmed to leave roughly 150 ms of silence before onset (and 300 ms after offset) in an attempt to make targets start roughly at the same time across trials. The average length of target stimuli was 2.7 seconds, while the average length of masker stimuli was 3.6 seconds. Target and masker lists were manually organized on the basis of their duration, so that targets were always shorter than the 2-sentence maskers for any target list/masker list combination. The masking sentences are openly available on OSF.

In contrast to target stimuli, the masker stimuli were created for the purpose of this study. All English, French, and Tamil masking stimuli were recorded by a single trilingual woman to keep speaking characteristics of the masker relatively constant. She acquired all three languages roughly simultaneously. She first translated all English transcripts into Tamil and then used the iPhone Voice Memo application using the internal microphone, holding the iPhone 10-15cm away from her mouth, in a quiet room in her home. She read each sentence from a script in her natural speaking voice with two seconds between each production. Recordings were broken down into 8 lists of 10 sentences, and she was instructed to leave one minute of silence at the start of a recording, which was subsequently used to filter out any background noise using a spectral subtraction method (Boll, 1979), conducted on Audacity version 2.1.1

(https://www.audacityteam.org/). An example masking sentence pair (English) is, “The small pup gnawed a whole in the sock. / The colt reared and threw the tall rider.” Audio files were cut in Audacity for disfluencies and extended pauses. The most fluent and natural productions were then selected, ensuring they contained few pauses between syllables (ideally continuously voiced). To create a masker list (of 10 maskers, each consisting of two simultaneous sentences spoken *in the same language*), five sentences were selected and added in pairs in all permutations. We chose to use a 2-talker masker because it is more difficult to ignore than a 1-talker masker but still sufficiently intelligible that they can attract listeners’ attention (see Hawley et al., 2004; Freyman et al., 2004). With 2-talker maskers, there are also fewer temporal dips (ie, gaps) in which participants can glimpse targets words compared to 1-talker maskers. We manually shifted in time each two-sentences pair to optimize the pseudo-stationarity of the combination waveform, in an attempt to leave as few temporal dips as possible (Collin & Lavandier, 2013; Leclère, Lavandier & Deroche, 2017). All maskers were finally root-mean-squared equalized at the same level as the targets (i.e., a target was as loud as a 2-voice masker, or roughly 3-dB louder than a single masking sentence), which defined a target-to-masker ratio (TMR) of 0 dB in this study.

3. Design and protocol

The competing speech task consisted of 20 blocks per participant, with 10 trials per block. Each participant began the study with two practice blocks, the first with English target sentences masked by English sentences, the second with French target sentences masked by Tamil sentences. None of the materials in the practice blocks were used in the rest of the experiment. Transcripts of the two masking sentences were displayed on the screen (depiction of experimental interface can be found in supplementary materials). Listeners were instructed not to

listen to them (as they were the female voices to ignore) and to listen instead to the third sentence spoken by a male voice.

The first trial of each block started with a target-to-masker ratio (TMR) at -16 dB, i.e., with a target sentence much quieter than the two maskers. Participants were given the opportunity to repeat the first trial as many times as necessary, with each repetition increasing the target level by 4 dB while the combined masker level was fixed. Participants were instructed to move on to the next trial once they were able to hear about 50% of the target sentence. At the end of each trial, participants were asked to type as much of the target sentence as they could. They were then presented with the correct transcript and asked to self-score the number of key words they correctly typed (see Appendix of Deroche, Limb, Chatterjee & Gracco, 2017b for a description of how accurate participants are at self-scoring in such tasks). Each target sentence contained five keywords, written in capital letters. If the listener identified three or more keywords correctly, the target level decreased by 2 dB, making the next trial more difficult. If the listener identified two or fewer keywords correctly, the target level increased by 2 dB, making the next trial easier. At the end of each block, this 1-up/1-down adaptive threshold method (Plomp & Mimpen, 1979) provided one speech reception threshold (SRT) calculated as the mean TMR over the last eight trials; it was assumed to bracket the TMR required to achieve 50% intelligibility.

After completing two practice blocks, participants completed 12 blocks measuring two SRTs for each of the six speech-on-speech situations (two target languages by three masker languages). While each of the target sentences was presented to every listener in the same order, the order of the masking conditions was rotated for successive listeners, to counterbalance effects of order and material. They then completed six blocks measuring three SRTs for each of the two

target languages against speech-shaped noise, where no transcript was displayed on the screen. Once again, these six blocks were counterbalanced.

4. Equipment

Because the experiment was delivered online during the COVID-19 pandemic, we were unable to control the audio quality presented to each participant. Instead, we asked participants to report whether they were listening through earbuds, headphones, loudspeakers, or through the default output of their computer. The two groups differed in the type of audio output ($\chi^2(3) = 12.3$, $p = .006$). In the L1 ENGLISH group, the most common audio output was the default output of their computer (36.8% of the group), while in the L1 FRENCH group, the most common audio output was headphones (52.9% of the group). This difference was likely negligible since the two groups did not differ from one another in their SRTs against either noise or speech maskers (see results). We also asked them to report on a scale of 1-5 how good their audio quality was, where 1 was "poor" and 5 was "excellent". The two groups did not differ in these subjective ratings ($\chi^2(2) = 2.9$, $p = .232$). Here again, we found no impact of audio quality on SRT performance with either noise or speech maskers (all p-values $\geq .252$). Participants were instructed to set the volume of their output to a comfortable level during the two practice blocks at the beginning of the task, and to not touch the volume afterwards. All stimuli were presented at a sampling frequency of 44.1 kHz, with a 32-bit resolution. All subjects provided informed consent online in accordance with the Institutional Review Board at Concordia University (ref: 30013650) and were compensated £7.50 for completing the study, or £3.75 in the case of withdrawal from the study.

3. Data analysis

1. Speech-on-noise conditions

The effect of target language was first examined from the SRTs collected against speech-shaped noise maskers. A linear mixed-effect model was fitted on the DV (being SRT in noise) with two fixed factors: group (L1 ENGLISH and L1 FRENCH) and target language (L1 and L2). We included random intercepts and slopes by participants and by lists. Each main effect and each interaction was tested by likelihood ratio tests progressively adding fixed terms to the final formula: DV ~ target*group + (1+target | participant) + (1+target | list).

2. Speech-on-speech conditions

A linear mixed-effect model was fitted on the SRT obtained across the six speech-on-speech conditions: with group (L1 ENGLISH and L1 FRENCH), target language (L1 and L2), and masker language (L1, L2, and Lf as fixed factors). We considered similar random terms as earlier, namely random intercepts and slopes (for the effect of target language) by participants and by lists. Furthermore, we also considered by-participant random slopes for the effect of masker (which improved the final model slightly further) while the model complexity could not support by-list random slopes for the effect of masker. Each main effect and each interaction was tested by likelihood ratio tests progressively adding fixed terms to the final formula: DV ~ target*masker*group + (1+target+masker | participant) + (1+target | list).

3. Correlational approach

Following on different calls from the bilingualism field to move beyond distinct categories and embrace the heterogeneity of bilinguals (Baum & Titone, 2014; Luk, 2015; de Bruin, 2019; DeLuca et al., 2019), we conducted additional analyses replacing group (L1 ENGLISH vs L1 FRENCH) by continuous variables. We made several attempts based on 1) age

at acquisition, 2) proficiency, and 3) L2 use, and we related the metrics to the listeners' performance as a function of target language (averaged over maskers) or as a function of masker language (averaged over targets). Perhaps surprisingly (see sections 4.3 and 4.4), these analyses suggested that performance varied categorically with respect to the target language: a small dominance in one language or another shifted the pattern of SRT quite dramatically. Thus, we finally reverted to a categorical approach to examine balanced versus unbalanced bilinguals and explored the parameter space by using different bilingualism metrics (AOA, listening use/proficiency, and speaking use/proficiency) and by changing the strictness of our balanced vs. unbalanced bilingualism definition within each different variable to confirm that our conclusions were not tied to a particular definition of a balanced bilingual based on AOA, proficiency, or use in their L2 (see Kremin & Byers-Heinlein, 2021, for a discussion of categorical vs. continuous approaches to bilingualism).

4. Results

1. Participants

For the most part, the two groups did not differ in demographics. The two groups were matched in student status (43.4% students; $\chi^2(2) = 0.2$, $p = .916$) and employment status (53.1% employed; $\chi^2(2) = 1.5$, $p = .462$) (see supplementary materials; employment status not shown). Note that some participants were missing data for their student and employment status leading to a third level (unknown status) in these two analyses.

The L1 ENGLISH and L1 FRENCH groups differed in country of residence ($\chi^2(3) = 97.0$, $p < .001$), which was expected as the countries from which participants were recruited have different official languages (English in the UK and USA, French in France, and both English and

French in Canada). Similar statistics were found with country of birth ($\chi^2(11) = 98.1$, $p < .001$) and nationality ($\chi^2(3) = 99.1$, $p < .001$).

Unintentionally, participants in the two groups differed in sex distribution ($\chi^2(1) = 6.0$, $p = .014$), as well chronological age ($t(126) = 3.0$, $p = .003$). The L1 ENGLISH group had a majority (70%) of female participants and was on average (std) 31.7 (9.5) years old, while the L1 FRENCH group was more balanced in sex (49% female) and was on average (std) 27.3 (6.9) years old. This likely had a negligible effect: be it for noise maskers (section 3.1) or speech maskers (section 3.2), sex did not interact with the factor of interest (group, target language, or masker language, all p-values ≥ .085). As for chronological age, it is well known that performance in speech perception tasks can degrade with age (e.g., Murphy, Daneman & Schneider, 2006; Schneider, Daneman & Pichora-Fuller, 2002; Schneider, Li & Daneman, 2007) but these effects are not expected until much later in life (often >60 years of age, e.g., Schneider, Speranza & Pichora-Fuller, 1998; Bilodeau-Mercure, Lortie, Sato, Guitton & Tremblay, 2015).

Unexpectedly, the two groups also differed in the amount of time they took to complete the study ($t(126) = -2.1$, $p = .037$). The L1 ENGLISH group took between 44.0 and 163.4 minutes, with a mean (std) of 73.7 (21.4) minutes, while the L1 FRENCH group took between 28.5 and 136.7 minutes, with a mean (std) of 82.1 (22.9) minutes. We do not think this is a very interesting observation: people generally wrote more words in their L1 than their L2, and written French is generally longer than written English. These two factors may have interacted to result in a longer completion time in the L1 FRENCH group.

Most importantly, the two groups differed (as intended) in their language background. All participants spoke at least French and English, with varying degrees of fluency. 17.2% of participants spoke three or four languages. Additional languages were Bulgarian, Chinese,

German, Italian, Luxembourgish, Moroccan Arabic (Darija), Russian, Spanish, and Welsh, with an average (std) listening proficiency of 6.3 (2.7) and an average (std) speaking proficiency of 5.1 (2.4). These L3 and L4 data were ignored in this study. Mixed analysis of variance (ANOVA) was conducted for each bilingualism metric with one between-subjects factor (group) and one within-subjects factor (language: L1 or L2). Results are illustrated in supplementary materials. The main effects and interactions are reported in Table 1. The main effect of language was always significant, confirming that all participants acquired their L1 much earlier than their L2 (0.9 vs 9.3 years old) and had better listening proficiency (9.9 vs 7.9), speaking proficiency (9.9 vs 7.3), listening use (9.8 vs 6.3), and speaking use (9.8 vs 5.2) in their L1 compared to their L2. None of this is surprising, but it gives a sense of the imbalance of this online sample of French-English bilinguals between their two languages. Less expected, the main effect of group and its interaction with language was significant for every variable except AOA). Post-hoc pairwise comparisons between the two groups were never significant in L1 (i.e., the fluency of the L1 FRENCH group in French was comparable to that of the L1 ENGLISH group in English) but were always significant in L2, namely that the L1 FRENCH group had better proficiency and greater use in English than the L1 ENGLISH group in French (mean difference (MD) for listening proficiency = 0.9, MD for speaking proficiency = 0.7, MD for listening use = 2.4, MD for speaking use = 1.7). This was not intended and may reflect the higher global use of English compared to French.

We expected these bilingualism measures to be highly correlated with one another. This was the case among all L2 proficiency and use variables (all $p < .001$, all $R^2 \geq .195$), but none of them correlated with L2 AOA (all $p \geq .298$, all $R^2 \leq .01$).

Table 1

Main effects and interactions of language and group

	Age of acquisition (AOA)	Listening proficiency	Speaking proficiency	Listening use	Speaking use
Main effect of language	$F(1, 126) = 395.2$, $p < .001$	$F(1, 126) = 298.3$, $p < .001$	$F(1, 126) = 336.3$, $p < .001$	$F(1, 126) = 255.8$, $p < .001$	$F(1, 126) = 319.6$, $p < .001$
Main effect of group	$F(1, 126) = 2.7$, $p = .105$	$F(1, 126) = 15.9$, $p < .001$	$F(1, 126) = 6.3$, $p = .014$	$F(1, 126) = 21.8$, $p < .001$	$F(1, 126) = 8.2$, $p = .005$
Interaction effect	$F(1, 126) = 1.3$, $p = .259$	$F(1, 126) = 11.8$, $p < .001$	$F(1, 126) = 4.7$, $p = .032$	$F(1, 126) = 35.5$, $p < .001$	$F(1, 126) = 15.5$, $p < .001$
L1 ENGLISH vs L1 FRENCH in L1	n/a	$t(126) = -0.4$, $p = .672$	$t(126) = -0.2$, $p = .822$	$t(126) = 0.6$, $p = .533$	$t(126) = 0.7$, $p = .456$
L1 ENGLISH vs L1 FRENCH in L2	n/a	$t(126) = -5.3$, $p < .001$	$t(126) = -3.3$, $p = .002$	$t(126) = -7.5$, $p < .001$	$t(126) = -4.8$, $p < .001$

Note: p-values less than .05 were considered significant. Holm correction applied to *t*-tests.

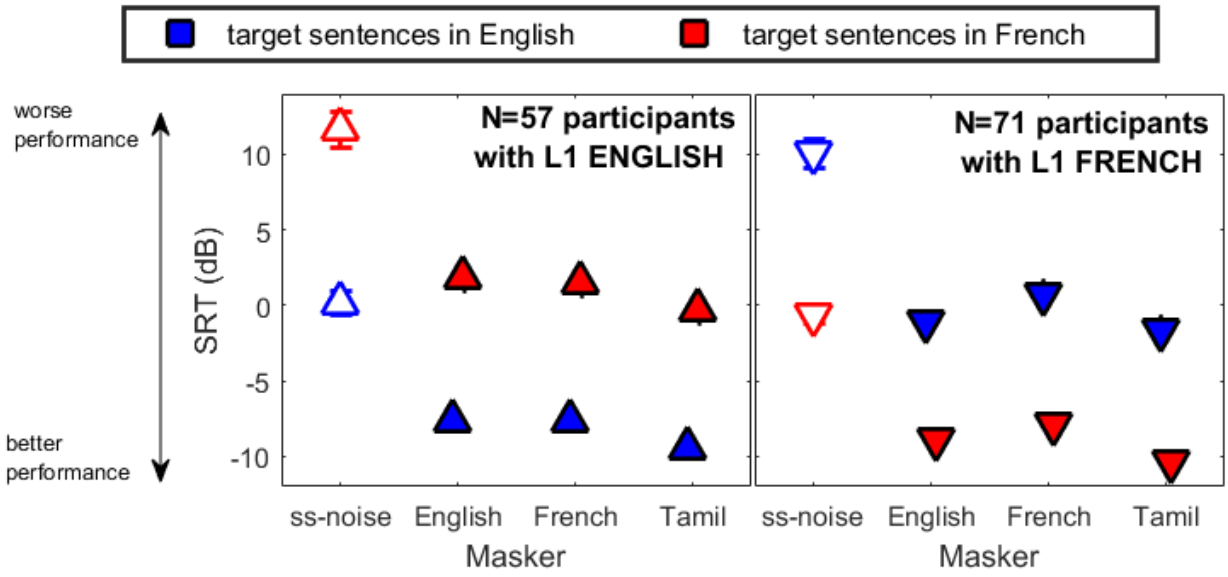
2. Speech-on-noise conditions

The LME revealed a main effect of target language ($\chi 2(1) = 26.4$, $p < .001$) reflecting that SRTs were estimated at 11.1 dB lower when listening to L1 rather than L2 (as illustrated in the left-hand sides of both panels in Fig. 1). There was no main effect of group ($\chi 2(1) = 1.0$, $p = .311$) and no interaction ($\chi 2(1) = 0.1$, $p = .749$). Participants performed better with L1 targets than with L2 targets, and this pattern was found equally in both groups. Given that participants in the L1 FRENCH group reported being more fluent in English relative to the L1 ENGLISH group

in French (see section 2.1 and supplementary materials), one could have expected a smaller SRT difference in L1 vs L2 in the L1 FRENCH than in the L1 ENGLISH group (i.e., an interaction), but this was not the case.

Figure 1

SRTs obtained across all experimental conditions



Note: L1 ENGLISH group depicted in the left panel, L1 FRENCH group depicted in the right panel. ss-noise is an abbreviation for speech-shaped noise, and the data points obtained with this masker type were empty symbols to differentiate them from the primary conditions of this study with speech maskers.

3. Speech-on-speech conditions

The linear mixed-effect model confirmed a main effect of target language ($\chi 2(1) = 48.5$, $p < .001$) reflecting that SRTs were estimated at 8.7 dB lower when listening to L1 rather than L2 (Fig. 1). There was also a main effect of masker language ($\chi 2(2) = 23.6$, $p < .001$), a key result, suggesting that SRT was respectively 0.7 and 2.3 dB lower with a L2 and a Lf masker compared to a L1 masker. Importantly, this masker effect did not interact with target language ($\chi 2(2) = 0.3$,

p = .846). To our knowledge this had never been demonstrated before. There was no main effect of group ($\chi 2(1) = 0.6$, p = .426) and group did not interact with target ($\chi 2(1) = 0.7$, p = .402), with masker ($\chi 2(2) = 1.7$, p = .437), or in a 3-way ($\chi 2(2) = 0.3$, p = .882). To summarize, in these speech-on-speech situations, participants found the task easier when attempting to listen to sentences spoken in L1 rather than spoken in L2, and this was true equally for L1 ENGLISH and L1 FRENCH participants (just like it was in background noise). On the other hand, all participants found it most challenging to ignore the female speakers conversing in their L1, somewhat intermediate when they conversed in their L2, and least challenging when they spoke a completely foreign language. Critically, this pattern disregarded whether the male target spoke in the participants' L1 or L2, whether it was French or English.

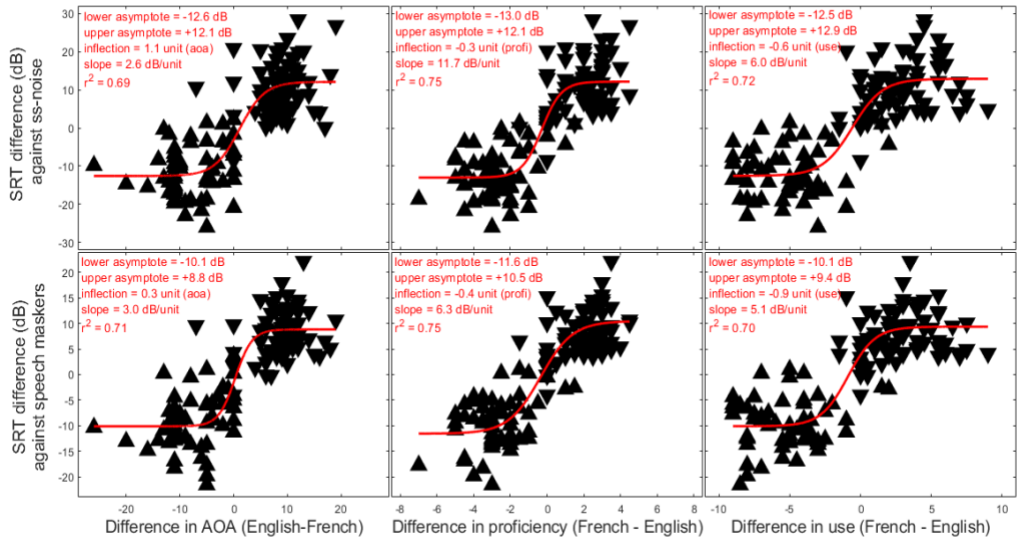
4. Effect of age at acquisition, proficiency, and use: target language

The results described above confirmed a strong effect of target language, namely participants had better speech understanding in their L1 than in their L2. This might seem to be an expected conclusion, but it becomes interesting when looking at individual data as a function of their proficiency differences between the two languages. Figure 2 illustrates the difference in SRT obtained in English vs French targets, with positive values indicating lower (better) threshold in French, and the abscissa always defined such that participants with better fluency in French are on the right side and participants with better fluency in English are on the left. Whether one considers this metric (the SRT difference) against noise maskers (top) or against speech maskers (bottom), it appeared to plateau as soon as participants exhibited a little dominance in one language or another. Here, we illustrated this dominance in three different ways: 1) from the fact that participants acquired their L1 earlier than their L2 (left panels), 2) from the fact that participants were more proficient in their L1 than in their L2 (middle panels),

and 3) from the fact that participants used their L1 more than their L2 (right panels). Note that we collected measures of *speaking proficiency* and *listening proficiency*, and similarly we collected measures of *speaking use* and *listening use*, as we had surmised that they would differ somewhat, but they did not. Figure 2 looked very similar with each one separately; so, for simplicity, the two measures of proficiency were averaged together (being highly correlated, r^2 = .64, $p < .001$) and the two measures of use were averaged together (being highly correlated, r^2 = .61, $p < 0.001$). Also note that the AOA, proficiency, and use for L1 were largely uninformative (being either at floor or ceiling), so the variability apparent in the abscissa of each panel is largely induced by L2.

Figure 2

Average SRT difference for French vs English targets



Note: Top panels show SRT difference against noise maskers. Bottom panels show SRT difference against speech maskers. SRT difference mapped as a function of three bilingualism-related metrics: difference in AOA (left), in proficiency (middle), and in use (right) between the two languages. Red curve depicts a sigmoid fit to the data. Participants in the L1 FRENCH group

(downward triangles) are shown with positive abscissa and (most likely) positive ordinate, while participants in the L1 ENGLISH group (upward triangles) are shown with negative abscissa and (most likely) negative ordinate.

A 4-parameter sigmoid function was fitted (with a non-linear least-square method) to these individual data in each panel, to try and better understand the impact of L1 dominance on the SRT difference observed between the two languages. It was of the form:

$$SRT\ diff = lower + \frac{(upper - lower)}{1 + exp^{\frac{infl - x}{s}}}$$

where ***lower*** and ***upper*** were, respectively, the lower and upper asymptote in dB; ***infl*** was the point of inflection on the abscissa, expressed in units of either AOA, proficiency, or use; ***s*** was a parameter influencing the shape of the slope at the inflection point, and was converted in dB / unit to highlight the impressive rate at which the SRT difference would flip to the opposite sign as L1 dominance swapped from English to French; and finally ***x*** was the abscissa, respectively, the difference in AOA, proficiency, or use between the two languages.

Against noise maskers (Fig. 2, top), the asymptotes reached +/- 12 to 13 dB which is quite meaningful: the estimated SRT obtained for unbalanced bilinguals did not continue to rise (i.e., worsen) as the proficiency in L2 decreased. In this paradigm, what is likely to happen is that participants with very low L2 proficiency would perform poorly not due to the audibility of L2 targets, but instead because of poor vocabulary or grammar/syntax mastery. Under the adaptive staircase method, the TMR would continue to rise without ever finding a reversal. As a consequence, these participants would be excluded from the analysis on the basis that a SRT could not reliably be measured. In other words, these asymptotes highlight the limitations of SRT measurements themselves: they become unreliable when participants cannot decipher words that

are clearly audible. This is a fundamental constraint of applying SRT protocols to listeners with poor proficiency in the target language.

Against speech maskers (here averaged across L1, L2, and Lf maskers, Fig.5, bottom), the asymptotes are on the order of +/- 9 to 11 dB. It is not exactly clear why they are less extreme than in the presence of noise. Perhaps, as listeners benefit from mechanisms such as dip-listening and F0 differences (Hawley et al., 2004; Collin & Lavandier, 2013; Deroche & Culling, 2013; Leclère et al., 2017), they can compensate to some degree for their lack of L2 proficiency, toning down the aforementioned issue.

Inflection points were generally close to 0 (within 1 year of AOA, or 1 unit in proficiency or use). To clarify, had the L1 FRENCH and L1 ENGLISH groups been exactly mirror images of each other, the inflection point should have occurred exactly at 0. The fact that it is slightly off is presumably a reflection that participants in the L1 FRENCH group had slightly better L2 than participants in the L1 ENGLISH group (see Demographics and supplementary materials).

Finally, and most importantly, the slopes of these fits were impressively steep, suggesting that there is a narrow range of AOA, proficiency, and use within which *balanced bilinguals* show a unique pattern of performance in this task. The middle-top panel provides the clearest demonstration of this fine sensitivity: compared to a fully balanced bilingual, a participant with a L2 proficiency of just one unit lower (e.g., at 9 rather than 10) would experience a 11.7 dB increase in SRT for L2 targets relative to L1 targets. Furthermore, as this value approaches the asymptote, such a participant would actually not appear drastically different (with regard to their performance) from a participant with much lower L2 proficiency. In other words, this demonstrates that a bilingualism-related metric such as L2 proficiency acts largely like a *categorical variable* for performance in speech-on-speech recognition. Note that for the other

panels (proficiency or use), the slopes were more modest, about 5 to 6 dB per unit suggesting a +/- 2 unit range where performance in this task might vary more gradually. And in the case of AOA, the slopes were shallower (2.6 to 3 dB per year) suggesting that it would take a difference of 4 or 5 years between the AOA of the two languages before a participant becomes a very unbalanced bilingual.

To summarize, this analysis delved deeper into the effect of L1 dominance and revealed two main findings. First, this task was ill-equipped to deal with participants with very low L2 proficiency. Though L2 proficiency was an eligibility criteria for participation, participants may not have realized the degree of L2 fluency necessary in L2 speech perception tasks. Second, the relative fluency between the two languages *acted like a categorical variable* to a large extent. Only a pocket of individuals who acquired their L2 within 4-5 years after their L1, and who had L2 proficiency and use within 1 or 2 units of their L1, showed some gradation in their performance across the two languages.

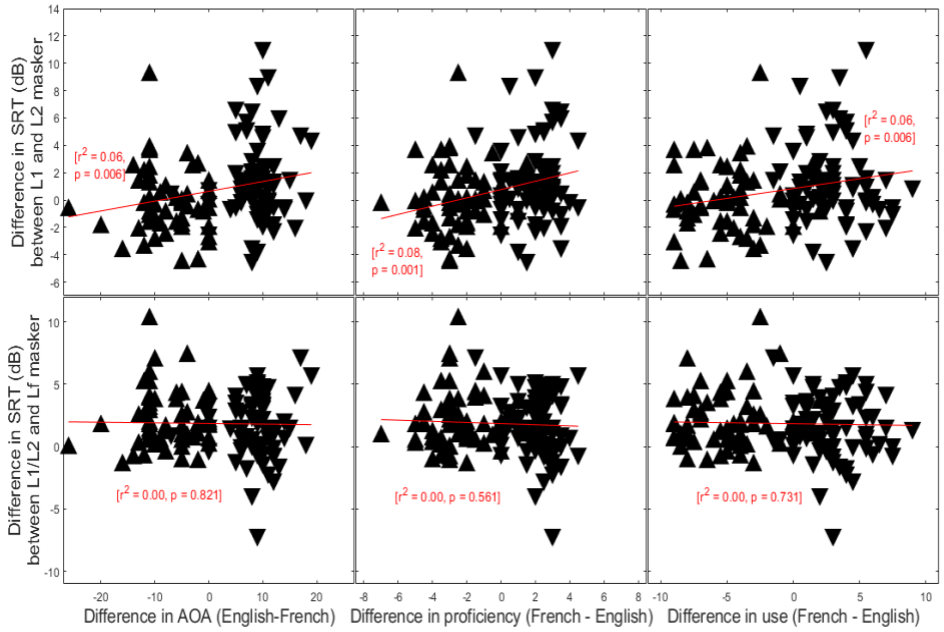
5. Effect of age at acquisition, proficiency, and use: masker language

The same approach was attempted to further delve into the effect of masker language, but it did not provide as useful insights. SRTs were averaged between the two target languages, and they were compared between L1 and L2 masker (Fig. 3, top) or between the average of L1 and L2 and Lf masker (Fig. 3, bottom). There was no longer any indication that the data followed a sigmoidal trend. L1 FRENCH participants (on the right side of the top panels) experienced more difficulty with French maskers and L1 ENGLISH participants (on the left side of the top panels) experienced more difficulty with English maskers. Given their location along the abscissa, this resulted in significant linear regressions ($p < .006$). But the large overlap between the two groups and the fact that there could be positive and negative ordinates within each group demonstrate

that masking effects are far from systematic. In the same vein, there was nothing striking across individuals' AOA, proficiency, and use, that made them more immune to Tamil maskers ($p >$ .561; Fig. 3, bottom). Thus, we opted for a different approach to further explore the effect of masker language (see next section).

Figure 3

Average SRT difference between language maskers



Note: Difference between the average SRT obtained with L1 vs L2 masker (top) and for L1/L2 vs Lf masker (bottom), averaged across the two target languages. Here, positive ordinates indicate French being a more effective masker than English (top) or French/English being more effective than Tamil (bottom).

6. Balanced versus unbalanced bilinguals

None of the previous analyses showed a main effect of group or an interaction between group and the effect of target or masker languages. For this reason, we could ignore whether participants were from the L1 FRENCH or L1 ENGLISH group and redefine them as balanced

versus unbalanced bilinguals, based on a cutoff of +/- 1 unit in proficiency. There was a total of 27 individuals whose L2 proficiency was within 1 unit of their L1 proficiency and formed the "balanced bilinguals" group. They were contrasted to 101 individuals who formed the "unbalanced bilinguals" group. A similar LME analysis was conducted as in section 4.2 and revealed an interesting dichotomy.

The newly defined groups resulted in a main effect ($\chi 2(1) = 5.1$, $p = .023$) and interacted with the target ($\chi 2(1) = 14.9$, $p < .001$), but not with the masker ($\chi 2(2) = 1.9$, $p = .381$) and not in a 3-way ($\chi 2(2) = 2.4$, $p = .301$). When listening to L1 targets, balanced bilinguals exhibited a small deficit (SRT on average 1.3 dB higher – green symbols in Fig. 4) relative to unbalanced bilinguals. In contrast, when listening to L2 targets, they exhibited an advantage (SRT on average 3.2 dB lower – purple symbols in Fig. 4) relative to unbalanced bilinguals. This pattern alone summarizes the pros and cons of bilingualism in this task: it may set bilinguals at a slight disadvantage in their L1 but is more than compensated by the benefit obtained when the task occurs in their L2.

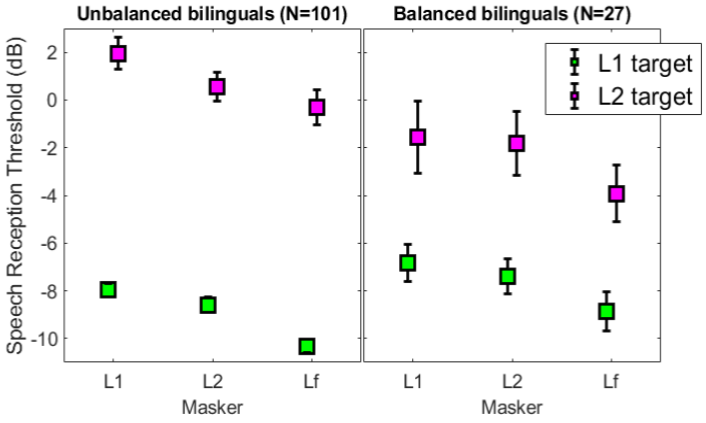
On the other hand, the lack of group by masker interaction here suggests that balanced and unbalanced bilinguals are (both) most affected by maskers speaking in L1, slightly less affected by maskers speaking in L2, and least affected by maskers speaking in a completely foreign language. This finding counters our suspicion that balanced and unbalanced bilinguals would be affected differently by different-language maskers.

Finally, note that these conclusions were robust to modifications of the cutoff used to allocate a participant as a balanced or unbalanced bilingual. We reiterated this analysis with a +/- 2 unit difference in proficiency (resulting in 58 vs 70 individuals, and the group by target interaction was $p = .003$), with a +/- 2 unit difference in use (resulting in 34 vs 94 individuals,

and the group by target interaction was $p = .021$), and with a 4 year difference in AOA (resulting in 20 versus 108 individuals, and the group by target interaction was $p < .001$). In all these attempts, the group by masker interaction or the 3-way interaction never reached significance ($p > .155$).

Figure 4

Comparison of unbalanced and balanced bilinguals



Note: Averaged speech reception threshold measured against speech maskers for unbalanced and balanced bilinguals defined according to their relative fluency in L1/L2. In contrast to Fig. 1, the targets are now defined relative to L1 and L2, irrespective of whether they were in English or in French. Error bars show one standard error from the mean, hence are larger in the restricted sample size of balanced bilinguals.

5. General discussion

1. Summary of findings

Returning to the bilingual cocktail party, we asked how different types of bilingual listeners would fare in different listening situations. Overall, we found that bilinguals performed best when a target spoke their L1 compared to their L2, and best when background speakers

spoke a foreign language, followed by their L2, followed by their L1. In other words, speech perception was at its best when L1 was the target, but impaired when L1 was the masking speech that had to be ignored. The difference between performance when listening to L1 and to L2 was strongest for unbalanced bilinguals, and more moderate for balanced bilinguals who performed somewhat less well in their L1 but better in their L2 compared to unbalanced bilinguals.

2. Balanced bilingual performance regarding target language

Balanced bilinguals performed relatively worse with L1 targets and relatively better with L2 targets compared to unbalanced bilinguals, and the relative advantage in L2 was larger than the relative disadvantage in L1. This finding is reminiscent of previous research which has found a bilingual deficit in understanding a target speaker in a multi-talker background. Regalado et al. (2019) found that early (balanced) bilinguals performed 2.7 dB better with L2 targets compared to late (unbalanced) bilinguals, which is comparable to the 3.2 dB balanced bilingual advantage we found (see section 4.5). Krizman, Bradlow, Lam and Kraus (2017) compared monolingual English and bilingual Spanish-English (L1 Spanish) speakers' performance on a task that used English targets, meaning that monolingual participants were being tested in their L1 while bilingual participants were being tested in their L2. We cannot directly compare our findings to those of Krizman et al. (2017) because there is a mismatch in target language type (L1 or L2) between their two groups. Monolinguals were unsurprisingly at an advantage in their design; however, they also found a bilingual *advantage* in tone-in-noise tasks. If language is considered as a spectrum of comprehension ranging from native speech to unintelligible sounds (such as those heard in the tone-in-noise tasks), then a non-native language would fall somewhere in between these two extremes. In this way, our results complement theirs, and show that

experimenters do not need to go as far as an artificial tone detection task to highlight a bilingual advantage effect; all it needs is a target that is not intrinsically biased in favor of monolinguals.

3. Masking release

All participants exhibited the weakest interference with foreign-language maskers, and the greatest interference with L1 maskers. Similar findings have been reported with monolingual samples (Rhebergen, Versfeld & Dreschler, 2005; Calandruccio, Brouwer, Van Engen, Dhar & Bradlow, 2013; Calandruccio, Leibold & Buss, 2016). Our estimate of a 2.3 dB difference in performance between L1 and Lf maskers is comparable to Rhebergen et al. (2005) who reported a 3.0 dB difference and Calandruccio et al. (2016) who reported a 2.8 dB difference in adults and a 3.0 dB difference in children for SRT against L1 or Lf maskers. In order to compare children to adults, this latter study used sentences from the Bamford-Kowal-Bench (BKB) Standard Sentence Test, which is based off of the speech of children aged 8-15. The fact that similar effect sizes were found with very different materials (BKB database vs IEEE database used in this experiment) and different age groups is a solid indication that the additional masking caused by the presence of a native language masker is a reliable and replicable phenomenon.

Our findings also agree, though less directly, with Calandruccio et al. (2013) and Lecumberri and Cooke (2006). In both of these studies, results were reported in terms of percentage of keywords that participants entered correctly, not SRT. In the first one, monolingual English participants were tested on English targets under three different masking conditions: English, Dutch, and Mandarin. Both Dutch and Mandarin were foreign languages, but Dutch is phonetically and semantically similar to English, in contrast to Mandarin. Listeners obtained approximately 20% increase in performance between Dutch and English maskers, and another 14% increase for Mandarin maskers. Considering that the slope of the psychometric functions

underlying performance in such tasks is generally around 10% per dB in the vicinity of the inflection point (Deroche et al., 2017b), this translates to a roughly 2.0 dB and 3.4 dB decrease in SRT between L1 maskers and Dutch and Mandarin maskers, respectively. This finding is not only in agreement with ours; it also suggests that some foreign languages are more foreign than others. Languages that are more phonetically and/or semantically different from L1 act as weaker maskers. Though this was not the goal of our study, we can use this finding to question how foreign Tamil was compared to French or English here. The lack of masker by target interaction, and the lack of 3-way interaction, suggests that it is similarly distinct from English and French, though such assertion would need to be formally examined with acoustic analyses (modulation spectrum, F0 profiles, intensity contours – which were conducted in the present materials but would not be sufficiently representative of these language as a whole). If we may rely on the 2.0 to 3.4 dB range suggested earlier, one might find that Tamil is not as foreign as Mandarin for listeners with either French or English as L1. Tamil belongs to a different language family (the Dravidian language family) than English and French. However, unlike Mandarin, it is not a tonal language, and so it might be more similar to English and French than Mandarin, though perhaps less similar than Dutch for instance which is an Indo-European language as are English and French.

Unlike Calandruccio et al. (2013), Lecumberri and Cooke (2006) used bilingual participants in their experiment. They compared performance of native English participants and native Spanish participants who spoke English as a second language in a task that involved identifying consonant phoneme sounds in a variety of noise conditions, including competing English and competing Spanish speech. The native English participants improved slightly by around 3% when the competing speech was Spanish (Lf) compared to English (L1), while native

Spanish participants barely improved (1% difference) when competing speech was English (L2) compared to Spanish (L1). Assuming the same 10% conversion between % to dB, this would equate to a minimal (0.1-0.3 dB) improvement between L1 and L2/Lf maskers. The most likely interpretation for this discrepancy is that their materials were less complex than ours. The ability to identify consonant phonemes would not be as affected by interfering speech as full sentences (like here), as less information needs to be identified and held in short-term memory with a simpler target. It is also possible that informational masking, though hardly quantifiable with speech maskers, would have a different impact (perhaps more concerned with attentional processes than linguistic competition) as a consonant target contains very little meaning compared to a full sentence target. Also note that performance was very high in their competing speech conditions (≥ 80%), making the % to dB conversion questionable.

Finally, there is evidence that the effects of L1 and Lf maskers can even be seen at a neurological level. Niemczak and Vander Werff (2021) conducted an electrophysiology study in which they compared auditory evoked potentials between target words in four masking conditions: English (L1), Dutch (Lf1), Mandarin (Lf2), and quiet (no masker). They found that in both high uncertainty and low uncertainty contents, the brain's ability to make sense of words (always spoken in L1) was more affected by L1 maskers than by Lf maskers. Though our behavioral findings cannot be directly compared with such neurophysiological metrics, it is very nice to start seeing the roots of such native language masking phenomena.

4. Continuous vs. categorical treatment of proficiency

One last aspect of the current findings that we should discuss in depth relates to the nature of variables that underlie the bilingual experience. Many papers published on the topic of bilingualism treat it as a categorical variable, dividing monolinguals from bilinguals (Bialystok,

Craik, Klein & Viswanathan, 2004; Rogers, Lister, Febo, Besing & Abrams, 2006; Filippi, Leech, Thomas, Green & Dick, 2012; Krizman et al., 2017; and Regalado et al., 2019, to name a few). However, several research groups have argued against dichotomizing continuous variables of bilingualism (Luk, 2015; de Bruin, 2019; DeLuca et al., 2019; Kremin & Byers-Heinlein, 2021). Indeed, doing so often oversimplifies data, reduces statistical power, and leads researchers to report conflicting results (MacCallum, Zhang, Preacher & Rucker, 2002). Although this is true of many other fields, regarding bilingualism more specifically, both Luk (2015) and de Bruin (2019) call to attention the fact that bilingualism is influenced by many factors: AOA, proficiency, and use, to external factors such as sociolinguistic context and demand involved in managing languages. An MRI study conducted by DeLuca et al. (2019) demonstrated the importance of considering multiple continuous facets of bilingualism by showing several correlations between connectivity networks in the brain and various "language experience" factors (AOA, language immersion, and language use).

While bilingualism may show itself to be a continuous and highly heterogeneous variable in other contexts, our data indicated that at least in a competing speech task, bilingualism might be better considered as a categorical rather than continuous variable, similar to other reports in speech segmentation (Cutler, Mehler, Norris & Segui, 1992). Our experimental design was certainly not set to arrive at this conclusion: in contrast to other studies comparing monolinguals and bilinguals, our sample did not have any monolingual participants, and included only bilinguals with varying L1/L2 fluency. Prepared for this heterogeneity, we also collected AOA, language proficiency, and language use, for listening or speaking, all of which in the hope that they would shine light on individual bilingual differences within the sample. However, even with the inclusion of various continuous measures of bilingualism, our data more strongly support a

language dominance effect where a slight imbalance between L1 and L2 results in dramatic SRT performance differences between the two languages, to the point that SRT protocol became ill-equipped. Put differently, we failed to see a continuous spectrum of performance that was correlated with the continuous measures of bilingualism captured through AOA, proficiency, and use. This is not to say that bilingualism should never be considered as a continuous variable, but rather that bilingualism might act as both a continuous and a categorical variable, depending on the nature of the study. Future bilingualism research should consider both continuous and categorical perspectives (Kremin and Byers-Heinlein, 2021) when analyzing their results to further contribute to our understanding of the nature of bilingualism in different contexts.

5. Future directions

Like many of the papers discussed above, our study looked at the performance of an adult participant pool. In the vein of Calandruccio et al. (2016), replicating our experimental design in children would be valuable both for developmental purposes (to better understand the role of bilingualism in language and cognitive development) and because children are generally known to be more prone to informational masking (e.g., Wightman, Kistler & O'Bryan, 2010) and therefore the native language masking phenomenon might be exacerbated in pediatric populations. This poses important challenges though, because this endeavor would require modifications of the task and materials to be more accessible to children (e.g., having a smaller vocabulary, shorter span of attention, and literacy skills) but doing so will precisely reduce the size of masking effects (which do not reveal their full magnitude until complex speech materials and tasks are used). Another avenue that could prove interesting would be replicating this experiment with bilinguals whose two languages are more distinct. Though English is a Germanic language and French is a Romantic language, the French language has had a large

influence on the English language as a result of French invasion of England in the 11th century (Britannica, 2021), and both are Indo-European. Perhaps we would find more dramatic results by recruiting bilinguals whose two languages were less related, such as English and Mandarin.

6. Conclusion

Our results indicate an overall effect of L1 dominance in speech perception contexts, both as a target and as a masker, regardless of whether individuals are balanced or unbalanced bilinguals. Though we did not see any overall balanced bilingual advantage in masking release, we did find a situational difference between balanced and unbalanced bilinguals, where balanced bilinguals showed a slight disadvantage with L1 targets compensated by a larger advantage with L2 targets. While there may not be a dramatic advantage or disadvantage to being bilingual, this situational difference in performance is incredibly sensitive to slight differences in perceptions about one's own language abilities, highlighting how minute changes in language background impact our speech understanding. Our findings suggest that we must take a flexible approach to studying bilingualism and be open to considering bilingualism under different modalities — both continuous and categorical — to fully understand how bilingualism shapes our communication abilities in multi-lingual cocktail party environments which are becoming common place.

Competing interests

The authors declare none.

Data availability

The data that support the findings are openly available in OSF at https://osf.io/2x653/?view_only=107b3ffbf42742a9964ddbf91e96765b.

References

Allen, JB (1995) How do humans process and recognize speech?. In *Modern methods of speech processing*. Boston, MA: Springer US, pp. 251-275.

Aubanel, V, Bayard, C, Strauß, A and Schwartz, JL. (2020) The Fharvard corpus: A phonemically-balanced French sentence resource for audiology and intelligibility research. *Speech Communication* 124, 68–74. https://doi.org/10.1016/j.specom.2020.07.004

Baum, S and Titone, D (2014) Moving toward a neuroplasticity view of bilingualism, executive control, and aging. *Applied Psycholinguistics* 35(5), 857–894. https://doi.org/10.1017/S0142716414000174

Bialystok, E, Craik, FIM, Klein, R and Viswanathan, M (2004) Bilingualism, Aging, and Cognitive Control: Evidence From the Simon Task. *Psychology and Aging* 19(2), 290. https://doi.org/10.1037/0882-7974.19.2.290

Bilodeau-Mercure, M, Lortie, CL, Sato, M, Guitton, MJ and Tremblay, P (2015) The neurobiology of speech perception decline in aging. *Brain Structure and Function* 220(2), 979–997. https://doi.org/10.1007/s00429-013-0695-3

Boll, S (1979) Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing* 27(2), 113-120.

Bregman, AS (1994) *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: The MIT press.

Britannica, The Editors of Encyclopaedia (2021, December 15). *Norman Conquest*. *Encyclopedia Britannica*. https://www.britannica.com/event/Norman-Conquest

Broersma, M and Scharenborg, O (2010) Native and non-native listeners' perception of English consonants in different types of noise. *Speech Communication* 52(11), 980–995. https://doi.org/10.1016/j.specom.2010.08.010

Bronkhorst, AW (2015) The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics* 77(5), 1465-1487.

Brungart, DS (2001) Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America* 109(3), 1101–1109. https://doi.org/10.1121/1.1345696

Brungart, DS, Simpson, BD, Ericson, MA and Scott, KR (2001) Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America* 110(5), 2527–2538. https://doi.org/10.1121/1.1408946

Calandruccio, L, Brouwer, S, Van Engen, KJ, Dhar, S and Bradlow, AR (2013) Masking Release Due to Linguistic and Phonetic Dissimilarity Between the Target and Masker Speech. *American Journal of Audiology* 22(1), 157–164. https://doi.org/10.1044/1059-0889(2013/12-0072)

Calandruccio, L, Dhar, S and Bradlow, AR (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. https://doi.org/10.1121/1.3458857

Calandruccio, L, Leibold, LJ and Buss, E (2016) Linguistic Masking Release in School-Age Children and Adults. *American Journal of Audiology* 25(1), 34–40. https://doi.org/10.1044/2015_AJA-15-0053

Calandruccio, L and 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 57(3), 1089–1097. https://doi.org/10.1044/2013_JSLHR-H-12-0378

Collin, B and Lavandier, M (2013) Binaural speech intelligibility in rooms with variations in spatial location of sources and modulation depth of noise interferers. *The Journal of the Acoustical Society of America* 134(2), 1146–1159. https://doi.org/10.1121/1.4812248

Cutler, A, Mehler, J, Norris, D and Segui, J (1992) The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology* 24(3), 381–410. https://doi.org/10.1016/0010-0285(92)90012-Q

de Bruin, A (2019) Not All Bilinguals Are the Same: A Call for More Detailed Assessments and Descriptions of Bilingual Experiences. *Behavioral Sciences* 9(3), 33. https://doi.org/10.3390/bs9030033

Declerck, M, Thoma, AM, Koch, I and Philipp, AM (2015) Highly proficient bilinguals implement inhibition: Evidence from n-2 language repetition costs. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 41(6), 1911–1916. https://doi.org/10.1037/xlm0000138

DeLuca, V, Rothman, J, Bialystok, E and Pliatsikas, C (2019) Redefining bilingualism as a spectrum of experiences that differentially affects brain structure and function. *Proceedings of the National Academy of Sciences* 116(15), 7565–7574. https://doi.org/10.1073/pnas.1811513116

Deroche, MLD and Culling, JF (2013) Voice segregation by difference in fundamental frequency: Effect of masker type. *The Journal of the Acoustical Society of America* 134(5), EL465–EL470. https://doi.org/10.1121/1.4826152

Deroche, MLD, Culling, JF, Lavandier, M and Gracco, VL (2017a) Reverberation limits the release from informational masking obtained in the harmonic and binaural domains. *Attention, Perception, & Psychophysics* 79(1), 363–379. https://doi.org/10.3758/s13414-016-1207-3

Deroche, MLD, Limb, CJ, Chatterjee, M and Gracco, VL (2017b) Similar abilities of musicians and non-musicians to segregate voices by fundamental frequency. *The Journal of the Acoustical Society of America* 142(4), 1739–1755. https://doi.org/10.1121/1.5005496

Filippi, R, Leech, R, Thomas, MSC, Green, DW and Dick, F (2012) A bilingual advantage in controlling language interference during sentence comprehension. *Bilingualism: Language and Cognition* 15(4), 858–872. https://doi.org/10.1017/S1366728911000708

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *The Journal of the Acoustical Society of America*, *115*(5), 2246–2256. https://doi.org/10.1121/1.1689343

Green, DW (1998) Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition* 1(2), 67–81. https://doi.org/10.1017/S1366728998000133

Grosjean, F (2012). Bilingualism: A short introduction. In Grosjean, F and Li, P (Eds.), *The Psycholinguistics of Bilingualism*. Hoboken, NJ: John Wiley & Sons, pp. 5–25.

Hawley, ML, Litovsky, RY, and Culling, JF (2004) The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America* 115(2), 833–843. https://doi.org/10.1121/1.1639908

Institute of Electrical and Electronics Engineers (1969) IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics* 17(3), 225-246.

Kidd, G, Mason, CR and Gallun, FJ (2005) Combining energetic and informational masking for speech identification. *The Journal of the Acoustical Society of America* 118(2), 982–992. https://doi.org/10.1121/1.1953167

Kidd, G., Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2008). Informational Masking. In Yost, WA, Popper, AN and Fay RR (Eds.), *Auditory Perception of Sound Sources* (Vol. 29). Boston, MA: Springer US, pp. 143-189. https://doi.org/10.1007/978-0-387-71305-2_6

Kilman, L, Zekveld, A, Hällgren, M and Rönnberg, J (2014) The influence of non-native language proficiency on speech perception performance. *Frontiers in Psychology* 5. https://doi.org/10.3389/fpsyg.2014.00651

Kremin, LV and Byers-Heinlein, K (2021). Why not both? Rethinking categorical and continuous approaches to bilingualism. *International Journal of Bilingualism* 25(6), 1560–1575. https://doi.org/10.1177/13670069211031986

Krizman, J, Bradlow, AR, Lam, SSY, and Kraus, N (2017) How bilinguals listen in noise: Linguistic and non-linguistic factors. *Bilingualism: Language and Cognition* 20(4), 834–843. https://doi.org/10.1017/S1366728916000444

Leclère, T, Lavandier, M and Deroche, MLD (2017) The intelligibility of speech in a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location. *Hearing Research* 350, 1–10. https://doi.org/10.1016/j.heares.2017.03.012

Lecumberri, MLG and 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. https://doi.org/10.1121/1.2180210

Linck, JA, Hoshino, N and Kroll, JF (2008) Cross-language lexical processes and inhibitory control. *The Mental Lexicon* 3(3), 349–374. https://doi.org/10.1075/ml.3.3.06lin

Linck, JA, Kroll, JF and Sunderman, G (2009) Losing Access to the Native Language While Immersed in a Second Language: Evidence for the Role of Inhibition in Second-Language Learning. *Psychological Science* 20(12), 1507–1515. https://doi.org/10.1111/j.1467-9280.2009.02480.x

Luk, G (2015) Who are the bilinguals (and monolinguals)? *Bilingualism: Language and Cognition* 18(1), 35–36. https://doi.org/10.1017/S1366728914000625

MacCallum, RC, Zhang, S, Preacher, KJ and Rucker, DD (2002) On the practice of dichotomization of quantitative variables. *Psychological Methods* 7(1), 19–40. https://doi.org/10.1037/1082-989X.7.1.19

Miller, G. A. (1947). The masking of speech. *Psychological Bulletin*, *44*(2), 105–129. https://doi.org/10.1037/h0055960

Murphy, DR, Daneman, M and Schneider, BA (2006) Why do older adults have difficulty following conversations? *Psychology and Aging* 21(1), 49–61. https://doi.org/10.1037/0882-7974.21.1.49

Niemczak, CE and Vander Werff, KR (2021) Informational Masking Effects of Similarity and Uncertainty on Early and Late Stages of Auditory Cortical Processing. *Ear & Hearing* 42(4), 1006–1023. https://doi.org/10.1097/AUD.0000000000000997

Oxenham, AJ, Fligor, BJ, Mason, CR and Kidd, G (2003) Informational masking and musical training. *The Journal of the Acoustical Society of America* 114(3), 1543–1549. https://doi.org/10.1121/1.1598197

Pivneva, I, Palmer, C and Titone, D (2012) Inhibitory Control and L2 Proficiency Modulate Bilingual Language Production: Evidence from Spontaneous Monologue and Dialogue Speech. *Frontiers in Psychology* 3. https://doi.org/10.3389/fpsyg.2012.00057

Plomp, R and Mimpen, AM (1979) Improving the Reliability of Testing the Speech Reception Threshold for Sentences. *International Journal of Audiology* 18(1), 43–52.

Qian, YM, Weng, C, Chang, XK, Wang, S and Yu, D (2018) Past review, current progress, and challenges ahead on the cocktail party problem. *Frontiers of Information Technology & Electronic Engineering* 19(1), 40-63.

Regalado, D, Kong, J, Buss, E and Calandruccio, L (2019) Effects of Language History on Sentence Recognition in Noise or Two-Talker Speech: Monolingual, Early Bilingual, and Late Bilingual Speakers of English. *American Journal of Audiology* 28(4), 935–946. https://doi.org/10.1044/2019_AJA-18-0194

Rhebergen, KS, Versfeld, NJ and Dreschler, WA (2005) Release from informational masking by time reversal of native and non-native interfering speech (L). *The Journal of the Acoustical Society of America* 118(3), 5.

Rogers, CL, Lister, JJ, Febo, DM, Besing, JM and Abrams, HB (2006) Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. *Applied Psycholinguistics* 27(3), 465–485. https://doi.org/10.1017/S014271640606036X

Schneider, BA, Daneman, M and Pichora-Fuller, MK (2002) Listening in aging adults: From discourse comprehension to psychoacoustics. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale* 56(3), 139–152. https://doi.org/10.1037/h0087392

Schneider, BA, Li, L and Daneman, M (2007) How Competing Speech Interferes with Speech Comprehension in Everyday Listening Situations. *Journal of the American Academy of Audiology* 18(07), 559–572. https://doi.org/10.3766/jaaa.18.7.4

Schneider, B. A., Daneman, M., & Pichora-Fuller, M. K. (2002). Listening in aging adults: From discourse comprehension to psychoacoustics. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *56*(3), 139–152. https://doi.org/10.1037/h0087392

Van Engen, KJ (2010) Similarity and familiarity: Second language sentence recognition in first- and second-language multi-talker babble. *Speech Communication* 52(11–12), 943–953. https://doi.org/10.1016/j.specom.2010.05.002

Van Engen, KJ and Bradlow, AR (2007) Sentence recognition in native- and foreign-language multi-talker background noise. *The Journal of the Acoustical Society of America* 121(1), 519–526. https://doi.org/10.1121/1.2400666

Warzybok, A, Brand, T, Wagener, KC and Kollmeier, B (2015) How much does language proficiency by non-native listeners influence speech audiometric tests in noise? *International Journal of Audiology* 54(sup2), 88–99. https://doi.org/10.3109/14992027.2015.1063715

Wightman, FL, Kistler, DJ and O'Bryan, A (2010) Individual differences and age effects in a dichotic informational masking paradigm. *The Journal of the Acoustical Society of America* 128(1), 11. https://doi.org/10.1121/1.3436536