Influence of social and semantic contexts in processing speech in noise.

Etienne Abassi & Robert J. Zatorre

Montreal Neurological Institute – McGill University,
3801 Rue University – Montréal, QC H3A 2B4 (Canada)

Centre for Research in Brain, Language and Music (CRBLM)
3640 de la Montagne, Montreal QC H3G 2A8 (Canada)

International Laboratory for Brain, Music and Sound Research (BRAMS)
90 Vincent-d'Indy Ave., Outremont, QC H2V 2S9 (Canada)

Correspondence to:
E.A. etienne.abassi@gmail.com
R.Z. robert.zatorre@mcgill.ca

Competing interests: All authors declare that they have no financial or non-financial competing interests.

ABSTRACT

Social interactions occupy a substantial part of our life. Not only interacting in first person, but also listening to others' interactions serve critical functions in understanding our social world. Yet, these auditory social signals are very often mixed with surrounding noise, such as in a restaurant, which degrades the information and forces our brain to use strategies to compensate for this loss. Although the role of semantics in speech comprehension under noisy conditions has been extensively studied as one of these strategies, the role of social context, and how it may interact with semantics, is unknown. Here, we conducted a series of four perceptual experiments to better understand the processing of multiple-speaker conversations from a third-person viewpoint, by manipulating the social and semantic contexts of a conversation. To do so, we used an assortment of artificial intelligence (AI) tools to generate a set of auditory stimuli, that consisted of two-speaker dialogues or one-speaker monologues (social context factor) arranged in intact or sentence-scrambled order (semantic context factor). Each stimulus comprised five sentences, with the fifth sentence embedded in multi-talker babbling noise. This fifth sentence was subsequently repeated without noise, with a single word altered or unchanged. Stimuli were presented over headphones to healthy young adult listeners, who were asked whether the repeated sentence was same as or different from the previous in-noise sentence. As results, we found overall that both the manipulation of social and semantic contexts had significant effects on performance in this task. We also found a correlation between a measure of autistic traits and individual performances when processing dialogues, but not monologues. Taken together, our results suggest that both semantic and social features of a conversation can modulate speech processing and that individual performance in this task could be related to certain social traits. These results raise new questions regarding the predictive or other mechanisms that may be at play when perceiving speech in social contexts.

INTRODUCTION

During our lifetime, we humans have to manifest complex abilities for navigating our social world (Adolphs, 2001, 2009). Consequently, a significant portion of our existence is devoted to engaging in social interactions with others. Not only interacting in first person, but also listening to others' interactions serves critical functions in understanding this world. Indeed, information about our social world comes first through our senses, and humans rely heavily on audition to communicate with their conspecifics (Adolphs, 2001). Thus, impairments in auditory functions can have detrimental effects on quality of life for a person (Ciorba et al., 2012; Dalton et al., 2003; Nordvik et al., 2018), by degrading their social interactions and provoking isolation from others. Therefore, a deeper understanding of the factors having an influence on perceiving social signals through audition is crucial both for the basic science of auditory cognition and for understanding the consequences of hearing loss.

A recurring complaint among those with hearing loss is the experience of sometimes being able to hear a conversation yet struggling to comprehend the spoken words, especially in acoustically challenging environments (Wilson, 2003). The importance of evaluating recognition performance of speech presented in background noise as an ecological measure of hearing loss was highlighted over 50 years ago (Carhart & Tillman, 1970). Despite this early recognition, it is only in the last decades that such evaluations have been integrated as a standard component of audiological test batteries (Strom, 2006). The conventional approach to assessing hearing loss – based primarily on pure-tone thresholds – has been found to offer limited predictive value for speech intelligibility in noise (Nilsson et al., 1994), prompting the development of novel assessments that specifically measure the ability to discern speech sounds from background noise across different signal-to-noise ratios (SNRs). Emerging from these efforts are tests such as the Hearing-in-Noise Task (HINT; Nilsson et al., 1994), the Speech-in-Noise Task (QuickSIN; Killion et al., 2004), and the Word-in-Noise Task (WIN; Wilson, 2003). While these assessments have significantly contributed to understanding hearing abilities in controlled environments, their applicability to real-life social situations, such as listening to a conversation between two or more speakers, remains limited. Indeed, these tests are all based exclusively on the recognition of words or sentences presented in isolation, yet in the context of a real-life conversation speech is never heard in an artificially isolated manner.

Traditionally, the assessment of word or sentence recognition in noise has heavily relied on bottom-up cues (Parthasarathy et al., 2020), manipulated through varying SNRs. However, recent research suggests that speech perception relies on additional, top-down processes. Indeed, contextual factors such as semantic cues have been shown to exert a substantial influence on speech perception, especially in challenging auditory environments. For instance, adding congruent or incongruent cues before a sentence (Zekveld et al., 2011, 2012, 2013), using word priming (Golestani et al., 2009, 2013), tweaking word probabilities (Roland et al., 2012) or using nonsense sentences (O'Neill et al., 2021) are some manipulations that showed a significance influence on speech perception. As for the mechanisms involved, it has been proposed that high-level semantic information would propagate down through hierarchical levels, first activating representations associated with the phonetic features of the speech and then with the lower-level acoustic features (Broderick et al., 2019; Sohoglu et al., 2012). In this framework, higher-level inferences would preactivate lower-level representations before the arrival of bottom-up auditory information (Kuperberg & Jaeger, 2016), that would facilitate the processing of bottom-up input.

Alternatively, it has been proposed that cognitive resources may be dynamically allocated between the semantic and auditory systems to deal with the perceptual and cognitive loads of incoming words (Mattys et al., 2009; Strauß et al., 2013). In this framework, words that are more dissimilar to their previous context would bear more information and thus would lead to a higher processing load on the semantic system (Broderick et al., 2019). As a result, a loaded semantic system would lead to reduced tracking of words in the auditory system. Critically, beyond these mechanisms it is important to note that, although we know speech patterns at different timescale are processed using different mechanisms and/or brain regions (DeWitt & Rauschecker, 2012; Farbood et al., 2015), all those prior behavioral studies only explored the influence of semantic context using isolated words or sentences, and they never manipulated semantic cues as part of a broader conversation-level context, that is a first gap we aimed to fill in the current study.

Going further, the influence of context in perception of a stimulus is not only known in the auditory modality but also in the visual modality (Brandman & Peelen, 2017; Oliva & Torralba, 2007), where recent work showed that a social context can have a significant influence on the perception of an individual visual stimulus and its neural processing (Abassi et al., 2020; Bellot et al., 2021). Although the hierarchy of information processing shows analogies in the visual and

auditory modalities (Iamshchinina et al., 2022), a significant gap exists in the literature on audition regarding the understanding of the impact of a social context on perceiving verbal sentences, especially in conditions where the auditory environment is challenging. The social context encompasses not only semantic cues but also involves intricate mechanisms related to the interplay between two or more speakers, elements that are generally overlooked in the examination of higher-order top-down cues, and that is a second gap we aimed to fill in the current study.

The present study seeks to delve into the influence of both semantic and social contexts in the understanding of conversations between two speakers in challenging listening conditions. More specifically, we first address the influence of semantic context by manipulating the global content of a conversation (i.e., intact vs. scrambled sentences drawn from different conversation) while keeping the content of the individual sentences unaltered. Here, if there is an effect of semantic context at the broad conversation level, we expect better processing of the intact conversations as compared to the sentence-scrambled ones. We then address the influence of social context by manipulating the speakers in a conversation (i.e., monologues vs. dialogues) while keeping the content of the conversations unaltered. Here, if there is an effect of social context at the broad conversation level, we expect better processing of the dialogues as compared to the monologues. Finally, we test whether individual differences in processing conversations can be linked to individual social traits. Here, we expect to find a significant relationship between individual performance in processing a conversation and an individual measure of autistic traits that are linked to sensitivity to social cues (AQ; Baron-Cohen et al., 2001).

METHODS

Participants. A total of 295 participants (for which details are given below) took part in the experiments online and were recruited on the prolific platform (*www.prolific.com*). Participants were remunerated 6£ for 40 minutes of their time. They were all native American English speakers with normal hearing (confirmed by self-reported online), aged between 18 and 35 years. This study was approved by the McGill University Research Ethics Board.

Stimuli. The entire set of stimuli can be found online at www.zlab.mcgill.ca/stimuli/. The main stimuli set consisted of English two-speaker dialogues or one-speaker monologues (factor: social context) arranged in intact or sentence-scrambled order (factor: semantic context) (Fig. 1). All stimuli were created using a combination of ChatGPT (GPT-3.5, *OpenAI*) to generate the contents

of the conversations and Google text-to-speech AI to convert the textual conversations into natural-sounding speech. More specifically, we first used ChatGPT to generate the text of 35 five-sentence conversations based on five themes (Music, Cinema, Sport, Literature and Travel). An example of the ChatGPT prompts used is *"On the topic of music, write X conversations of five sentences each between two persons with a maximum of fifteen words per sentences"*. We did not use the exact conversation contents as generated by ChatGPT, but we made modifications such as word, syntax and length changes. For each five-sentence conversation, we created a sixth sentence, that was same as the fifth sentence, but with only one word change that could appear at the beginning, the middle or the end of the sentences. This last sentence was the only one presented in noise (see below).

Prior to converting the conversations to natural-sounding speech, we added prosody tags in each sentence to make it sound more naturalistic, following the Speech Synthesis Markup Language (SSML) rules. An example of an edited sentence with prosody tags is *"*I *<prosody rate="fast" pitch="+4st">* love *</prosody>* playing *<prosody pitch="-2st">* music *</prosody>"*. We then selected two voices from the list proposed by Google text-to-speech AI: one female voice (*'en-US-Wavenet-G'*) and one male voice (*'en-US-Wavenet-J'*). We finally converted each text sentence of each conversation into auditory speech for the two voices, and we normalized all audio tracks to have the same rms amplitude across all sentences.

We then arranged the sentences into conversations as follows. For intact dialogues, half of the conversations began with the female voice and alternated with the male voice, and for the other half it was the opposite. For monologues, half of the conversations were composed of the female voice only, while the other half was composed of the male voice only. Furthermore, for both dialogues and monologues, sentence-scrambled conversations were created by randomly shifting sentences across the conversations, except for the first and fifth sentences. In total, we created 240 conversational stimuli (70 stimuli by unique condition, half primed with female voice and half primed with male voice).

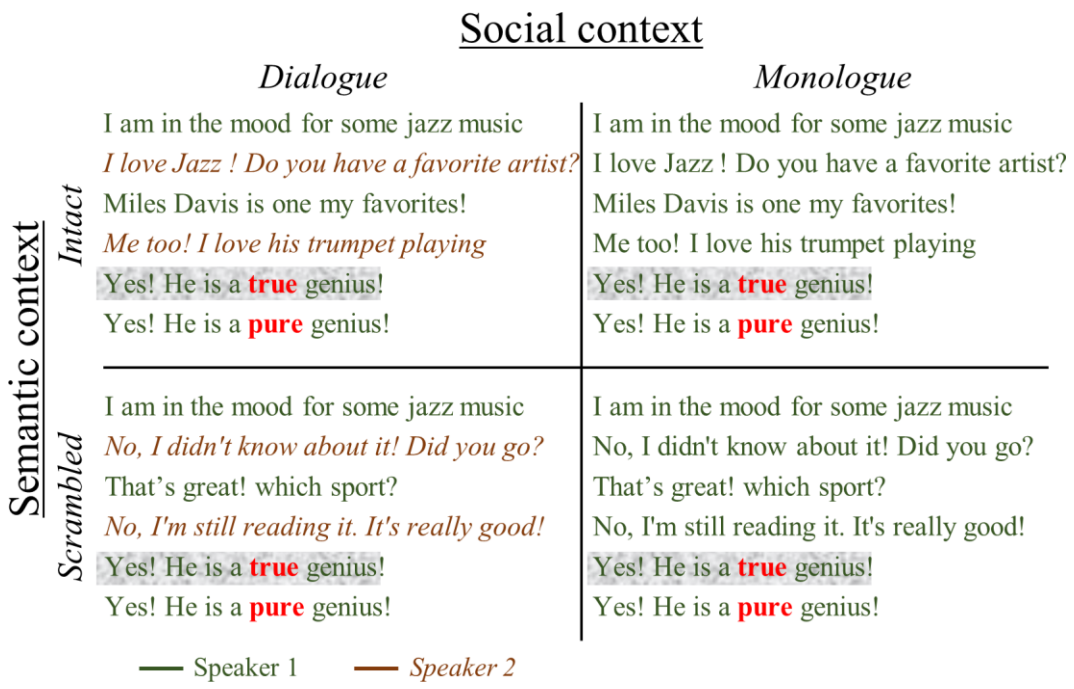


Figure 1. Example of conversations composing the stimuli set. Stimuli are arranged following a 2x2 design with factors social context (Dialogue / Monologue) and semantic context (Intact / Scrambled). For dialogues, sentences presented in normal green alternating with brown italic represents the alternance between the two speakers (one female and one male. Sentences presented over grey background illustrate that the fifth sentence of each conversation was always embedded in noise. The last sentence could be either the same or different than the previous one, with only one word changing for the different sentences, as represented in bold red.

In addition to dialogues and monologues, we created a baseline set of 50 one-sentence stimuli that were different than the sentences used in the conversations, that were used as a baseline condition. For each of these sentences, we created another sentence, that was same as the original sentence, but with only one word change that could appear at the beginning, the middle or the end of the sentences. The word change itself was unrelated to the content of the conversation—that is, it was a neutral word within the sentence that would be equally appropriate in the context (see Fig 1). We then converted these sentences into natural-sounding speech for both female and male voices, using the same method than for conversations. More specifically, one example of the ChatGPT prompts used to generate the text of these one-sentence stimuli was *"Write X sentences on the topic of music with a maximum of fifteen words per sentences".* We then added prosody tags following SSML rules and transformed it into audio using Google text-to-speech. In total, we created 100 one-sentence stimuli (50 for female and 50 for male).

Multi-talker babbling noise. In a second step, we embedded the fifth sentence of each conversation and the one-sentence stimuli into multi-talker babbling noise. We used multi-talker babbling noise as a masker for our stimuli because we wanted it to be as naturalistic as possible (for a review on the different type of maskers, see Assmann & Summerfield, 2006). The rationale to create this noise was as follow: 1) We used a total of 8 competing voices (4 females and 4 males), that is a good compromise between creating sufficient interference with the target voices while keeping a reasonable amount of competing voices (Cooke, 2006; Li & Lutman, 2006; Miller, 1947; Simpson & Cooke, 2005); 2) We selected the fundamental frequencies (f0) of the competing voices so it would not overlap with the target voices, as having too close f0 between target and competing voices may interfere with the task (Assmann, 1999; Bird, 1997; Brokx & Nooteboom, 1982; David et al., 2017; Lee & Humes, 2012); 3) We selected a f0 range between ~121 Hz and ~201 Hz by steps of around one semi-tone between each voice (see supp. Fig. 1), so the range would be large enough to cover a representative range of f0 in the general population (see Oliveira et al., 2021).

To generate the content of the noise sentences, we again used ChatGPT to generate 24 semantically anomalous sentences in English. Anomalous sentences were used here to eliminate the possibility that participants might extract an entire meaningful sentence from a speaker other than the target while keeping a correct grammar (Van Engen & Bradlow, 2007). An example of the ChatGPT prompts used is *"Write X anomalous sentences"* and an example of the anomalous sentences generated is *"Her purple umbrella whistled a dolphin"*. We then selected eight voices from the list proposed by Google text-to-speech AI: 4 females voices (*'en-US-News-K'; 'en-US-News-L'; 'en-US-Wavenet-E'; 'en-US-Wavenet-C'*) and 4 male voices (*'en-US-Wavenet-D'; 'en-US-News-N'; 'en-US-Wavenet-I'; 'en-US-News-M'*). Prior to generating the audio speech for the noise, we set the f0 parameter for each voice as follows. For all voices we used the vowel /ɑ/ as a reference (Cristina Oliveira et al., 2021; De Krom, 1994), then for each voice we measured the original f0 of the voice for the duration of this vowel using the built-in function 'pitch' of Matlab (R2022b; The MathWorks Inc.) and time windows of 52 ms., and we defined the f0 as the median of the f0 measured for all time windows. We then calculated by how much we should increase or decrease the original f0 of the voice to get to the custom f0 we wanted.

After generating the 24 audio sentences for each of the 8 voices, we created the multi-talker noise stimuli using custom Matlab codes, closely following the method of Van Engen & Bradlow (2007): 1) Each audio track was normalized to have the same rms amplitude across all tracks; 2) For each talker, two randomly chosen sentences were concatenated to ensure the duration of the noise tracks would exceed the durations of all target sentences; 3) A random multiple of 100 ms (0–500 ms) of silence was added at the beginning of each concatenated sentences to stagger the talkers once they were mixed together; 4) All eight talkers were mixed, by randomly choosing one concatenated sentence by talker, and the initial 500 ms of the concatenated sentences was removed to eliminate noise that did not contain all eight talkers; 5) The first 100 ms of the completed noise tracks was faded in. In total, we produced 120 multi-talker noise stimuli. In a final step, we mixed the fifth sentences of each of the previously generated conversations or each of the one-sentence stimuli with a randomly chosen noise stimulus for the entire duration of the sentences, to produce SNRs of -1, -4, and −6 dB.

Procedure. All experiments were developed using PsychoPy (Peirce et al., 2019), then transformed into JavaScript using the PsychoJS library, as implemented in PsychoPy, and hosted on the Pavlovia platform (www.pavlovia.org) for online running of the experiments. Participants were then recruited on the Prolific platform (www.prolific.com) and the study was restricted only to people with a desktop computer or a laptop. Prior to enrolling in the study, online participants were instructed that wearing headphones or earbuds was mandatory to participate. They were also asked to be in a silent and calm environment, as well as remove any object around them that may disturb their attention, such as a smartphone or a TV.

Prior to beginning the study, participants went through a series of checks to ensure optimal ‘auditory hygiene’ while performing the tasks. To do so, we used guidelines and scripts from Zhao et al. (2022). More specifically, participants first heard multi-talker noise at a root-mean-squared amplitude that would produce SNRs of -1 dB if they had been presented in the real experiment and mixed with the stimulus sentences, and they were instructed to adjust their audio volume at a comfortable level, so they would not be exposed to overly loud sounds during the experiment. They then performed a headphone screening using Antiphase tone (Woods et al., 2017) to confirm that they were wearing headphones. More specifically, for this task participants had to correctly detect an acoustic target in an intensity-discrimination task. On each trial, they were presented with

three consecutive 200 Hz sinusoidal tones in random order, and they were asked to determine which of the tones was perceptually the softest. Two of the tones were presented diotically: 1) the ‘standard’, and 2) the ‘target’ that was presented at -6dB relative to the standard. The third tone (the ‘foil’) had the same amplitude as the standard but was presented dichotically such that the left and right signals had opposite polarity (anti-phase, 180°). Over headphones, the standard and foil should have the same loudness, making the target clearly distinguishable as softer. In contrast, over loudspeakers the left and right signals interact destructively before reaching the listener’s ears, resulting in a weaker acoustic signal at both ears, and thus a weaker loudness sensation for the foil than the target, causing participants to respond incorrectly. Following the headphone check, participants had to read and sign electronically the consent form, after which they were allowed to participate in the study.

Each experiment (for which specific differences will be given below) was preceded by a training task. For each training task, sentence-in-noise SNRs were of -1 dB for the first half of the stimuli and of -4 dB for the second half, while each training trial was followed by feedback. The experiment consisted of two main tasks: 1) a baseline task, in which only one-sentence stimulus were presented in noise, that was used to assess the level of understanding of sentences-in-noise for each participants without prior social or semantic context given and 2) a conversational task, in which dialogues or monologues in intact or sentence-scrambled order were presented, with the fifth sentences of each stimuli embedded in noise, that was used to assess the effects of prior social or semantic context given on the processing of sentences embedded in noise. For both tasks, immediately after hearing a sentence in noise, the participants heard it again but without noise. Half of the time, these repeated sentences were exactly the same as the one presented with noise, while the other half had a one-word difference. The task for the participant was to determine, as quickly as possible, whether the repeated sentence was exactly the same or different than the one presented in noise, using the left or right arrows on their keyboard.

A typical trial began with a waiting time of 500 ms with a fixation cross at the center of the screen, followed by the presentation of the auditory stimuli, then by a visual recall of the response buttons until the participants pressed one of the two buttons, and afterwards it began again. Every 32 trials, the participants were invited to take a break. The response buttons were counterbalanced across participants (i.e., half of the participants used the left arrow for the ‘same’

trials and right arrow for the ‘different’ trials, while the other half did the opposite). The first voice heard was also counterbalanced across participants (i.e., half heard the female voice first in the first presented stimuli, while the other half heard the male voice in first). For all tasks, the order of trials was pseudo-randomized to avoid the presentation of more than 3 time of the same trial type (same/different), target voice (male/female), social factor (dialogue/monologue), semantic factor (intact/scrambled) or SNR (-4/-6 dB; for Exp. 1 only).

At the end of the experiment, participants completed the 50-item Autism-Spectrum Quotient scale (AQ; Baron-Cohen et al., 2001), which measures the degree to which an adult with normal intelligence shows traits associated with the autism spectrum, in the domain of social skills, attention-switching, attention to details, communication, and imagination. Subjects responded to each item using a 4-point rating scale, ranging from “definitely agree” to “definitely disagree”. All responses (Accuracy, reaction time [RTs] and AQ questionnaire) were collected with Pavlovia and processed with Matlab.

Data preprocessing. For all experiments, the following data preprocessing steps were applied: 1) Participants were excluded from the study if their average Accuracy in the baseline task was less than 54%, meaning they were not able to do the experiment correctly, that could be due to various reasons (e.g., low quality headphones, low volume, noisy environment, inattention etc.); 2) For each participant, trials for which RTs were >3 SD above the participant mean were discarded for both RTs and Accuracy; 3) For each participant, trials in which the participant’s response was inaccurate were discarded for RTs only; 4) For each participant, both Accuracy and RTs mean performances for each condition were normalized using the baseline performance measured in the baseline task using this formula: [(Performance-Baseline)/Baseline]. This last step was done to take into account inter-subject variability in performing the task; 5) Participants were excluded from the study if their mean normalized RTs averaged across all conditions was >3 SD above the group mean.

Analyses. For all experiments, normalized mean Accuracy (mean proportion of correct responses) and normalized mean RTs were used to compare the effect of intact vs. scrambled conversation across dialogues and monologues. To do so, for experiment 1 we used a Mixed-Model ANOVA (between factor: social context; within factor: semantic context). For the three other experiments, we used repeated-measures ANOVAs with two factors: social context (Dialogue/Monologue) x

semantic context (Intact/Scrambled). Post-hoc analyses were performed using pairwise *t* tests (two-tailed).

EXPERIMENT 1

Goal. Experiment 1 was conducted to choose a correct SNR for measuring our effects of interest in the subsequent experiments. To do so, we selected two SNRs of -6 and -4 dB, that are known to produce a speech reception threshold (SRT, see Plomp & Mimpen, 1979) that is in the range of 40% to 60% depending on the studies and/or methods (Brungart et al., 2001; Dingemanse & Goedegebure, 2020; Festen & Plomp, 1990; Hagerman, 1984; McArdle et al., 2005; Miller et al., 1951).

Participants. Experiment 1 included two subgroups, with one group presented with dialogues only and one group presented with monologues only. Group 1 was composed of 37 participants (10 female, 2 sex undetermined, mean age 28.7 years +/- 5.92 SD) after exclusion of 3 participants. Group 2 was composed of 33 participants (13 female, 1 sex undetermined, mean age 30.9 years +/- 6.4 SD) after exclusion of 7 participants. Exclusion of participants was based on the criteria given in the method before.

Procedure. For the two groups, the experiment began with two training tasks: one with 8 baseline trials and the other with 4 conversation trials. The order of the training tasks was counterbalanced over participants. If a participant obtained less than 85% correct responses during the training part, (s)he had to redo the training to pursue the study. If a participant failed the training part for a second time, (s)he was not allowed to pursue the study and was prematurely excluded from the study. Following the training, participants underwent the experiment that consisted of two baseline blocks (32 trials each) and two conversation blocks (32 trials each) interleaved, for which the order was counterbalanced across participants. Importantly, for each condition, half of the trials used an SNR of -4 dB, while the other half used an SNR of -6 dB randomly assigned, with a different set of stimuli for each SNR.

Results. Prior to use the baseline to normalize the data, for the SNR of -6 dB we found a mean baseline accuracy of 61.9 % (+/- 1.9 SEM) for group 1 and 64.5 % (+/- 1.8 SEM) for group 2, while for the SNR of -4 dB, we found a mean baseline accuracy of 73.2 % (+/- 1.8 SEM) for group 1 and 72.3 % (+/- 1.8 SEM) for group 2. For the main task, for the SNR of -6 dB we found a mean

average accuracy of 61.7 % (+/- 1.8 SEM) for group 1 and 60.2 % (+/- 1.9 SEM) for group 2, while for the SNR of -4 dB we found a mean average accuracy of 78.1 % (+/- 1.8 SEM) for group 1 and 77.1 % (+/- 1.5 SEM) for group 2.

After using the baseline to normalize the data, for the SNR of -6 dB (fig. 2a), neither RTs nor Accuracy showed any effect at all of the two factors. The mixed-model ANOVA for RTs showed no effects of social ($F(1,68) = 2.07$, $p = 0.154$, $np2 = 0.03$) or semantic context ($F(1,68) = 2.56$, $p = 0.114$, $np2 = 0.04$), and no interaction between the two ($F(1,68) = 0.04$, $p > 0.250$, $np2 = 0.01$). The mixed-model ANOVA for accuracy showed no effects of social ($F(1,68) = 1.36$, $p = 0.248$, $np2 = 0.02$) or semantic context ($F(1,68) = 0.47$, $p > 0.250$, $np2 = 0.01$), and no interaction between the two ($F(1,68) = 0.76$, $p > 0.250$, $np2 = 0.01$).

For the SNR of -4 dB (fig. 2b), using RTs the mixed-model ANOVA showed no effect of social context ($F(1,68) = 0.55$, $p > 0.250$, $np2 = 0.01$), but a close-to-significant trend for an effect of semantic context ($F(1,68) = 3.81$, $p = 0.055$, $np2 = 0.05$). There was no interaction between the two factors ($F(1,68) = 2.20$, $p = 0.143$, $np2 = 0.03$). Post-hoc *t*-tests showed lower RTs for intact vs. sentence-scrambled conversations for monologues, but no differences for dialogues (Dialogues: $t(36) = 0.33$, $p > 0.250$; Monologues: $t(32) = 2.52$, $p = 0.017$). Using Accuracy, the mixed-model ANOVA showed no effects of social ($F(1,68) = 0.01$, $p > 0.250$, $np2 = 0.01$) or semantic context ($F(1,68) = 0.33$, $p > 0.250$, $np2 = 0.01$). However, interestingly, it showed a close-to-significant trend for an interaction between these two factors ($F(1,68) = 3.80$, $p = 0.055$, $np2 = 0.05$). Post-hoc *t*-tests showed higher Accuracy for intact vs. sentence-scrambled conversations for dialogues, but no differences for monologues (Dialogues: $t(36) = 2.10$, $p = 0.043$; Monologues: $t(32) = 0.84$, $p > 0.250$)

To summarize, we first found that an SNR of -6 dB was producing accuracies in the range of 60-65%, while and SNR of -4 dB was producing accuracies in the range of 72-78%, that was a first indication that an SNR of -6 dB might produce ceiling effect and thus was not adequate, while and SNR of -4 dB might not. In addition, we found that using an SNR of -6 dB showed no effects at all for either RTs or Accuracy, while using an SNR of -4 dB showed promising results for both RTs and Accuracy. Thus, we used an SNR of -4 dB for all subsequent experiments.

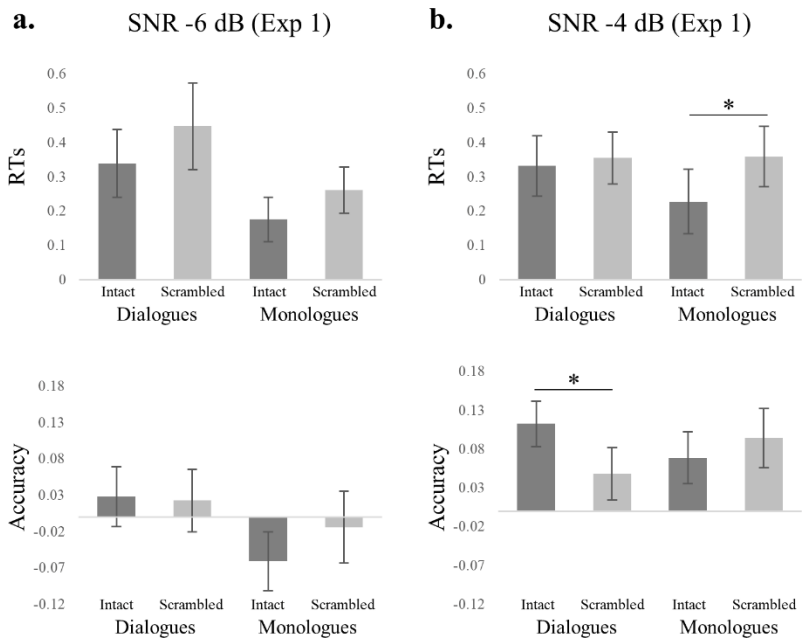


Figure 2. Behavioral data for experiment 1. Response times (RTs) and proportion of correct responses (Accuracy) for intact and scrambled dialogues or monologues with SNRs of (**a.**) -6 dB and (**b.**) -4 dB. *$p \leq 0.05$.

EXPERIMENT 2

Goal. Experiment 2 was specifically designed to address the effect of semantic context while controlling for the effect of social context. To do so, separate blocks of dialogues and blocks of monologues were presented to the same participants. Here, we expected to find a main effect of semantic context using the repeated measures ANOVA in either RTs or Accuracy.

Participants. Experiment 2 was composed of 70 participants (28 female, 2 sex undetermined, mean age 31.7 years +/- 5.83 SD) after exclusion of 4 participants based on the criteria given in the method before.

Procedure. Experiment 2 used an SNR of -4 dB and began with the baseline task. More specifically, at first a training task featuring 6 baseline trials was followed by a testing task of 44 baseline trials. If a participant failed more than 54% of the baseline trials, (s)he was excluded from the study. Afterwards, participants were presented with one block of dialogues only and one block of monologues only (32 trials each, with half intact and half sentence-scrambled conversations, presented in random order). The order of blocks was counterbalanced across participants. Each block was preceded by a training task of 4 trials of the same condition.

Results. Using RTs, the repeated measures ANOVA showed no effect of social context ($F(1,69) = 0.40$, $p > 0.250$, $np2 = 0.01$), but a main effect of semantic context ($F(1,69) = 10.74$, $p = 0.002$, $np2 = 0.13$), with no interaction between these two factors ($F(1,69) = 0.02$, $p > 0.250$, $np2 = 0.01$). Post-hoc *t*-tests (fig. 3a) showed lower RTs for intact vs. sentence-scrambled conversations for both dialogues and monologues (Dialogues: $t(69) = 2.68$, $p = 0.009$; Monologues: $t(69) = 2.38$, $p = 0.020$).

Using Accuracy, the repeated measures ANOVA showed no effects of social ($F(1,69) = 0.80$, $p > 0.250$, $np2 = 0.01$) or semantic context ($F(1,69) = 1.46$, $p = 0.231$, $np2 = 0.02$), and no interaction between the two ($F(1,69) = 0.02$, $p > 0.250$, $np2 = 0.01$).

EXPERIMENT 3

Goal. Experiment 3 was specifically designed to address the effect of social context while controlling for the effects of semantic context. To do so, separate blocks of intact conversations and blocks of sentence-scrambled conversations were presented to the same participants. Here, we expected to find a main effect of social context using the repeated measures ANOVA in either RTs or Accuracy.

Participants. Experiment 3 was composed of 59 participants (21 female, mean age 32.1 years +/- 5.81 SD) after exclusion of 7 participants based on the criteria given in the method before.

Procedure. Experiment 3 used an SNR of -4 dB and began with the baseline task. More specifically, at first a training task featuring 6 baseline trials was followed by a testing task of 44 baseline trials. If a participant failed more than 54% of the baseline trials, (s)he was excluded from the study. Afterwards, participants were presented with one block of intact conversations only and one block of sentence-scrambled conversation only (32 trials each, with half dialogues and half monologues presented in random order). The order of blocks was counterbalanced across participants. Each block was preceded by a training task of 4 trials of the same condition.

Results. Using RTs, the repeated measures ANOVA showed a trend for an effect of social context ($F(1,58) = 2.85$, $p = 0.097$, $np2 = 0.05$), but no effect of semantic context ($F(1,58) = 0.32$, $p > 0.250$, $np2 = 0.01$), and no interaction between these two factors ($F(1,58) = 0.22$, $p > 0.250$, $np2 = 0.01$). Post-hoc *t*-tests (fig. 3b) showed no differences of RTs for intact vs. sentence-scrambled

conversations for both dialogues and monologues (Dialogues: $t(58) = 0.85$, $p > 0.250$; Monologues: $t(58) = 0.16$, $p > 0.250$).

Using Accuracy, the repeated measures ANOVA showed no effects of social context ($F(1,58) = 0.01$, $p > 0.250$, $np2 = 0.01$) but an effect of semantic context ($F(1,58) = 8.62$, $p = 0.005$, $np2 = 0.13$), and no interaction between these two factors ($F(1,58) = 0.01$, $p > 0.250$, $np2 = 0.01$). Post-hoc *t*-tests showed higher Accuracy for intact vs. sentence-scrambled conversations for dialogues and a trend for this effect for monologues (Dialogues: $t(58) = 2.18$, $p = 0.034$; Monologues: $t(58) = 1.91$, $p = 0.062$).

EXPERIMENT 4

Goal. Experiment 4 was specifically designed to address both the effects of semantic and social context at the same time. To do so, mixed blocks of dialogues and monologues in intact or sentence-scrambled configurations were presented to the same participants. Here, we expected to find both effects of social and semantic contexts using the repeated measures ANOVA in either RTs or Accuracy.

Participants. Experiment 4 was composed of 68 participants (36 female, 1 undermined, mean age 29.7.1 years +/- 5.5 SD) after exclusion of 7 participants based on the criteria given in the method before.

Procedure. Experiment 4 used an SNR of -4 dB and began with two training tasks: one with 8 baseline trials and the other with 4 conversation trials. The order of the training tasks was counterbalanced over participants. If a participant did less than 85% correct responses during the training part, (s)he had to redo the training to pursue the study. If a participant failed the training part for a second time, (s)he was excluded from the study. Following the training, participants underwent the experiment that was constituted of two baseline blocks (32 trials each) and two conversation blocks (32 trials each) interleaved, for which the order was counterbalanced across participants.

Results. Using RTs, the repeated measures ANOVA showed effects of both social ($F(1,67) = 5.39$, $p = 0.023$, $np2 = 0.07$) and semantic context ($F(1,67)= 10.12$, $p = 0.002$, $np2 = 0.13$), but no interaction between these two factors ($F(1,67) = 1.17$, $p > 0.250$, $np2 = 0.02$). Post-hoc *t*-tests (fig. 3c) showed lower RTs for intact vs. sentence-scrambled conversations for dialogues only

(Dialogues: $t(67) = 3.03$, $p = 0.004$; Monologues: $t(67) = 1.11$, $p > 0.250$). Here, we conducted an additional one-way repeated measure ANOVA to compare the mean RTs across all four experimental conditions (Intact Dialogues, Scrambled Dialogues, Intact Monologues, Scrambled Monologues). We found that mean RTs were significantly different from each other ($F(3,268)= 4.95$, $p = 0.002$, $np2 = 0.05$). Critically, post-hoc *t*-tests showed that these differences were largely driven by the intact dialogues for which RTs were significantly lower than every other conditions (*vs.* scrambled dialogues: $t(67) = 3.03$, $p = 0.004$; *vs.* intact monologues: $t(67) = 3.06$, $p = 0.003$; *vs.* scrambled monologues: $t(67) = 4.44$, $p < 0.001$).

Using Accuracy, the repeated measures ANOVA showed no effects of social context ($F(1,67) = 1.94$, $p = 0.168$, $np2 = 0.03$) or semantic context ($F(1,67) = 1.56$, $p = 0.216$, $np2 = 0.02$), and no interaction between these two factors ($F(1,67) = 2.07$, $p = 0.155$, $np2 = 0.03$).

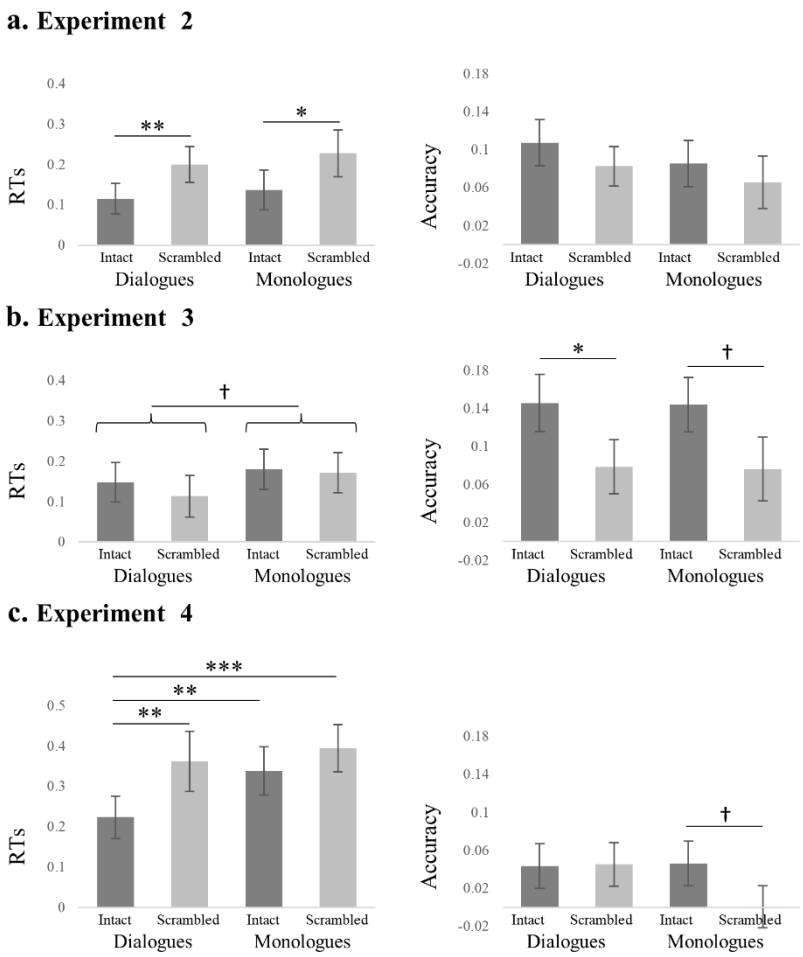


Figure 3. Behavioral data for experiments 2, 3 and 4. Response times (RTs) and proportion of correct responses (Accuracy) for intact and scrambled dialogues and monologues for experiments 2 (**a.**), 3 (**b.**) and 4 (**c.**). †p ≤ 0.10; *p ≤ 0.05; **p ≤ 0.01; ***p ≤ 0.001.

COMBINED EXPERIMENTS

Goal. Here, we combined the data of experiments 2, 3 and 4, that have the exact same conditions within the same participants, to increase the statistical power of our analyses.

Participants. The participants were the same as in experiment 2, 3 and 4, except that we excluded all participants for which RTs averaged across all conditions was >3 SD above the combined group mean. Thus, the combined experiment was composed of 193 participants after exclusion of 4 participants.

Results. Using RTs, the repeated measures ANOVA showed effects of both social ($F(1,192) = 8.52$, $p = 0.004$, $np2 = 0.04$) and semantic context ($F(1, 192)= 10.01$, $p = 0.002$, $np2 = 0.05$), but no interaction between these two factors ($F(1, 192) = 0.13$, $p > 0.250$, $np2 = 0.01$). Post-hoc *t*-tests (fig. 4a) showed lower RTs for intact vs. sentence-scrambled conversations for dialogues, and a trend for monologues (Dialogues: $t(192) = 2.37$, $p = 0.019$; Monologues: $t(192) = 2.50$, $p = 0.013$). Here, we conducted an additional one-way repeated measure ANOVA to compare the mean RTs across all four experimental conditions (Intact Dialogues, Scrambled Dialogues, Intact Monologues, Scrambled Monologues). We found that mean RTs were significantly different from each other ($F(3,768)= 6.97$, $p < 0.001$, $np2 = 0.03$). Critically, post-hoc *t*-tests showed that these differences were largely driven by the intact dialogues for which RTs were significantly lower than every other conditions (*vs.* scrambled dialogues: $t(192) = 2.37$, $p = 0.019$; *vs.* intact monologues: $t(192) = 2.27$, $p = 0.024$; *vs.* scrambled monologues: $t(192) = 4.45$, $p < 0.001$).

Using Accuracy, the repeated measures ANOVA showed no effect of social context ($F(1, 192) = 2.04$, $p = 0.155$, $np2 = 0.01$), but an effect of semantic context ($F(1, 192)= 11.17$, $p = 0.001$, $np2 = 0.05$), and no interaction between these two factors ($F(1, 192) = 0.26$, $p > 0.250$, $np2 = 0.01$). Post-hoc *t*-tests showed higher Accuracy for intact vs. sentence-scrambled conversations for both dialogues and monologues (Dialogues: $t(192) = 2.04$, $p = 0.042$; Monologues: $t(192) = 2.54$, $p = 0.012$).

To study whether we could find a relationship between performances in processing conversations and individual's social abilities, we performed a Spearman correlation between the individual AQ scores and the individual difference scores for intact minus scrambled conversations, separately for dialogues and monologues. Using RTs, for dialogues (fig. 4b) but not monologues (fig. 4c) we found that there was a negative correlation between AQ and processing of intact vs. scrambled conversations (Dialogues: *rho*(197) = -0.18, p = 0.010; Monologues: *rho*(197) = 0.07, $p > 0.250$). Furthermore, we found that these two correlations were significantly different from each other (p = 0.008) using a percentile bootstrap approach (Wilcox, 2016). We didn't find significant correlations using Accuracy (Dialogues: *rho*(197) = -0.03, $p > 0.250$; Monologues: *rho*(197) = 0.08, $p > 0.250$) with no differences between dialogues and monologues ($p < 0.250$).

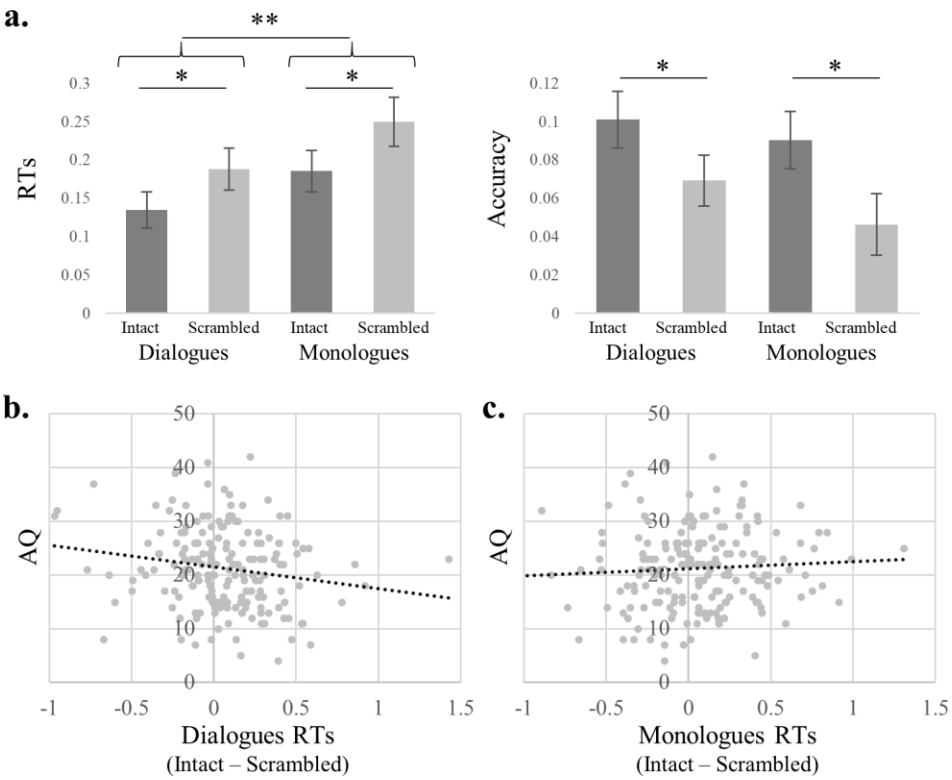


Figure 4. Behavioral data for combined experiments. a. Response times (RTs) and proportion of correct responses (Accuracy) for intact and scrambled dialogues and monologues for combined experiments (2, 3 and 4). *p ≤ 0.05; **p ≤ 0.01. **b. c.** Spearman correlation between the individual AQ scores and the individual difference scores for intact minus scrambled dialogues (b) and monologues (c).

DISCUSSION

The present study was designed to investigate the processing of multiple-speaker conversations from a third-person perspective through a set of four experiments. The first experiment was designed to choose the correct SNR to be used in subsequent experiments, that was set to -4 dB. The second experiment was designed to assess specifically the effect of semantic context, and showed this effect on response time only, but no effect of social context in either response time and accuracy. The third experiment was designed to assess specifically the effect of social context, and showed a trend for this effect on response time only, as well as a significant effect of semantic context in accuracy only. The fourth experiment was designed to assess both the effect of semantic and social context, and showed both effects on response times only. Combined data from experiment two, three and four showed effects of both semantic and social context on response times, and an effect of semantic context on accuracy only. They also showed that the processing of dialogues, but not monologues, was correlated with an individual score of social traits.

Our finding that semantically related information can facilitate the understanding of spoken speech in the presence of masking sound is in line with previous studies that showed similar effects using a variety of approaches (Golestani et al., 2009, 2013; O'Neill et al., 2021; Roland et al., 2012; Zekveld et al., 2011, 2012, 2013). However, even though speech patterns at different timescales are processed using different mechanisms and/or brain regions (DeWitt & Rauschecker, 2012; Farbood et al., 2015), none of these studies examined the effects of semantics in the context of an entire conversation; instead, they only examined the effects of semantics in isolated words or sentences. Here, we addressed this question using a more naturalistic approach at the larger timescale of a conversation. Thus, although we found that, similarly to isolated words or sentences, there is a significant effect of semantic context when processing sentences embedded in whole conversation, an open question is whether these effects are relying on the same mechanisms and/or brain regions than for shorter timescales. Overall, our finding is coherent with what one might expect within a predictive coding framework (Clark, 2013; Friston, 2005; Rao & Ballard, 1999). Indeed, within this framework, it is believed that accurate prior knowledge leads to a reduction in error signal between top-down predictions and bottom-up sensory information, which results in perceptual enhancement and anoverall change in neural activity (Sohoglu et al., 2012). Note

however that the effect we measured was not trivially due to detecting a semantic anomaly in the altered sentence, because the changed word always resulted in a semantically appropriate sentence. Thus, the predictive mechanism is operating at a more abstract level, perhaps related to sentence cohesion, than at a more local level of single-word error detection.

Furthermore, the fact that, under at least some circumstances, an individual is quicker and better at correctly processing a sentence in the context of a dialogues as compared to a monologue is in keeping with the idea that the human perceptual system is tuned to socially relevant stimuli (Belin et al., 2000; Issard et al., 2023; Levy et al., 2003). However, while previous research consistently showed that people are better at understanding dialogues than monologues (Branigan et al., 2011; Fox Tree, 1999; Tolins et al., 2018), this effect has been usually linked to the internal content of a dialogue as compared to a monologue, rather than to its intrinsic social value as an interaction. Indeed, dialogues usually contain information that monologues lack, such as discourses markers and/or different perspectives from multiple interlocutors, that may help in the understanding of the information that is shared (Pickering & Garrod, 2021). While this hypothesis may provide a partial explanation for this advantage in certain situations, it cannot apply to our study. Here, we used the exact same contents for both dialogues and monologues, with the only change being the sex of the speakers and their alternance or not. Furthermore, our task required the listener to identify if a word had been changed, but that word was not itself relevant to the content of the dialogue (e.g., see figure 1, in which the changed word was “true” vs “pure” both of which fit the meaning of the sentence, and neither of which is related to the dialogue content). Thus, we rather hypothesize that the presence of social cues, such as two speakers interacting with each other, would trigger a sharpening of the content of sentences by the perceptual system, that we observed for dialogues but not monologues. This hypothesis is in line with recent findings in the domain of visual perception showing that a human body is better perceived when facing another body (i.e., such as in a social interaction) as compared to facing in an opposite direction (Abassi & Papeo, 2023; Papeo et al., 2017; Papeo & Abassi, 2019; Vestner et al., 2019, 2021).

In addition, we found that the processing of dialogues, but not monologues, correlates with individual autism-spectrum traits, so that the magnitude of the difference in performances between intact and scrambled dialogues decreases as the AQ score increases. In the auditory modality, some correlations between perceptual performance for processing social stimuli and autistic traits has

been found for voice recognition (Gervais et al., 2004; Schelinski et al., 2016; Skuk et al., 2019) and phonetic categorization (Stewart & Ota, 2008), but wasn't yet found for the processing of sentences. In the visual modality, such correlations have been found for the processing of faces (Wyer et al., 2012) or biological motion (van Boxtel et al., 2017), but also for processing two bodies engaged in a prototypical interaction (Abassi & Papeo, 2022). The results of this last study strongly echo the results that we found in the current study, in the sense that they both show a correlation between perceptual processing of social interactions and individual social abilities, independently of the visual or auditory modality. This raise the question of whether the influence of social traits on processing social stimuli is linked to the perceptual modality in itself, or would rather be linked to a domain-general mechanisms that would be generally involved in processing social interactions, independently to the modality. While further experiments will be needed to approach this question, behavioral effects that reliably characterize the performance of the individual, highlighting population-level and interindividual differences (see also Pesquita et al., 2022), appear especially worthy of attention as they can offer handy tools for detection of diseases or vulnerabilities.

Overall, another point to highlight is that effects of social context were found only with RTs, while effects of semantic context were found in both RTs and accuracy. A first reason for this difference may be that our task could be more fine-tuned (e.g., by using an adaptative SNR; see Soleymani et al., 2019) to better take into account individual variabilities that would be reflected in the performances in the task and especially accuracy, that is more sensitive to floor or ceiling effects than RTs (Magnus et al., 2017; Willoughby et al., 2023). Another reason could be due to the nature of the measured effects themselves. Indeed, it has been proposed that RTs and accuracy may be affected by different attentional cues via distinct cognitive and/or neural mechanisms (Prinzmetal et al., 2005; van Ede et al., 2012; but see Mulder & Van Maanen, 2013). Thus, effects of social and semantic contexts may rely on different attentional cues, that may be reflected in measuring RTs and accuracy, while the exact nature of these cues, if exist, would need a more thorough exploration. A last point is that, while we found both effects of social and/or semantic contexts depending the design of our experiments, we never found an interaction between these two factors. Although this is not a limitation in itself, previous literature showed a strongest perceptual sensitivity for stimuli implying multiple interactors as compared to stimuli with individuals, at least in the visual system (Isik et al., 2017; Walbrin & Koldewyn, 2019). Thus, we

could have expected to find that the difference of performance in processing intact versus scrambled conversations would be higher for dialogues than monologues. Here, previous attempts in finding analogous effect in the brain has been so far unconclusive (Landsiedel & Koldewyn, 2023; Olson et al., 2023), so this interaction, if it exists, has yet to be found. Still, in the current study we found the lowest RTs in processing intact dialogues as compared to every other condition in some experiments. This last finding, alongside the presence of a correlation between processing dialogues and autistic traits, may represent clues for the potential existence of a special mechanism in processing canonical verbal social interactions between multiple interactors as compared to other forms.

By exploring the influence of social context on speech perception in naturalistic social scenarios, the present research contributes valuable insights into the intricate interplay between contextual factors (i.e., social and semantic context) and hearing abilities. Furthermore, in sharing our carefully designed set of stimuli, we hope to provide useful materials for further research and/or potential clinical applications. Overall, these findings hold the potential to inform not only our understanding of auditory perception for speech but also the development of interventions and technologies that better align with the complexities of real-world listening experiences for individuals with hearing loss.

Acknowledgments

This research was supported by a postdoctoral fellowship to EA and by the Grand Prix Scientifique to RJZ from the Fondation Pour l'Audition (FPA RD-2022-1). RJZ holds a Canada Research Chair and is a fellow of the Canadian Institute for Advanced Research.

References

Abassi, E., & Papeo, L. (2022). Behavioral and neural markers of visual configural processing in social scene perception. *NeuroImage*, 119506. https://doi.org/10.1016/j.neuroimage.2022.119506

Abassi, E., & Papeo, L. (2023). Category-selective representation of relationships in visual cortex. *Journal of Neuroscience*, JN-RM-0250-23. https://doi.org/10.1523/JNEUROSCI.0250-23.2023

Abassi, E., & Papeo, L. (2020). The representation of two-body shapes in the human visual cortex. *Journal of Neuroscience*, *40*(4), 852–863. https://doi.org/10.1523/JNEUROSCI.1378-19.2019

Adolphs, R. (2001). *The neurobiology of social cognition*. 231–239.

Adolphs, R. (2009). The social brain: Neural basis of social knowledge. *Annual Review of Psychology*, *60*, 693–716. https://doi.org/10.1146/annurev.psych.60.110707.163514

Assmann, P. F. (1999). Fundamental frequency and the intelligibility of competing voices. *Proceedings of the 14th International Congress of Phonetic Sciences*, 179–182.

Assmann, P., & Summerfield, Q. (2006). The Perception of Speech Under Adverse Conditions. *Speech Processing in the Auditory System*, 231–308. https://doi.org/10.1007/0-387-21575-1_5

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5–17. https://doi.org/10.1023/A:1005653411471

Belin, P., Zatorre, R. J., Lafallie, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature 2000 403:6767*, *403*(6767), 309–312. https://doi.org/10.1038/35002078

Bellot, E., Abassi, E., & Papeo, L. (2021). Moving Toward versus Away from Another: How Body Motion Direction Changes the Representation of Bodies and Actions in the Visual Cortex. *Cerebral Cortex*, *31*(5), 2670–2685. https://doi.org/10.1093/cercor/bhaa382

Bird, J., Darwin, C. J., Palmer, A. R., Rees, A., Summerfield, A. Q., & Meddis, R. (1998). Effects of a difference in fundamental frequency in separating two sentences. Psychophysical and physiological advances in hearing, 263-269.

Brandman, T., & Peelen, X. M. V. (2017). *Interaction between Scene and Object Processing Revealed by Human fMRI and MEG Decoding*. *37*(32), 7700–7710. https://doi.org/10.1523/JNEUROSCI.0582-17.2017

Branigan, H. P., Catchpole, C. M., & Pickering, M. J. (2011). What makes dialogues easy to understand? *Language and Cognitive Processes*, *26*(10), 1667–1686. https://doi.org/10.1080/01690965.2010.524765

Broderick, M. P., Anderson, A. J., & Lalor, E. C. (2019). Semantic Context Enhances the Early Auditory Encoding of Natural Speech. *Journal of Neuroscience*, *39*(38), 7564–7575. https://doi.org/10.1523/JNEUROSCI.0584-19.2019

Brokx, J. P. L., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, *10*(1), 23–36. https://doi.org/10.1016/S0095-4470(19)30909-X

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, *110*(5), 2527–2538. https://doi.org/10.1121/1.1408946

Carhart, R., & Tillman, T. W. (1970). Interaction of Competing Speech Signals With Hearing Losses. *Archives of Otolaryngology*, *91*(3), 273–279. https://doi.org/10.1001/ARCHOTOL.1970.00770040379010

Ciorba, A., Bianchini, C., Pelucchi, S., & Pastore, A. (2012). The impact of hearing loss on the quality of life of elderly adults. *Clinical Interventions in Aging*, *7*, 159–163. https://doi.org/10.2147/CIA.S26059

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, *119*(3), 1562–1573. https://doi.org/10.1121/1.2166600

Cristina Oliveira, R., Gama, A. C. C., & Magalhães, M. D. C. (2021). Fundamental Voice Frequency: Acoustic, Electroglottographic, and Accelerometer Measurement in Individuals With and Without Vocal Alteration. *Journal of Voice*, *35*(2), 174–180. https://doi.org/10.1016/J.JVOICE.2019.08.004

Dalton, D. S., Cruickshanks, K. J., Klein, B. E. K., Klein, R., Wiley, T. L., & Nondahl, D. M. (2003). The Impact of Hearing Loss on Quality of Life in Older Adults. *The Gerontologist*, *43*(5), 661–668. https://doi.org/10.1093/GERONT/43.5.661

David, M., Lavandier, M., Grimault, N., & Oxenham, A. J. (2017). Sequential stream segregation of voiced and unvoiced speech sounds based on fundamental frequency. *Hearing Research*, *344*, 235–243. https://doi.org/10.1016/J.HEARES.2016.11.016

De Krom, G. (1994). Consistency and Reliability of Voice Quality Ratings for Different Types of Speech Fragments. *Journal of Speech and Hearing Research*, *37*(5), 985–1000. https://doi.org/10.1044/JSHR.3705.985

Dingemanse, G., & Goedegebure, A. (2020). Efficient Adaptive Speech Reception Threshold Measurements Using Stochastic Approximation Algorithms. *Trends in Hearing*, *23*, 1–17. https://doi.org/10.1177/2331216520919199

Festen, J. M., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America*, *88*(4), 1725–1736. https://doi.org/10.1121/1.400247

Fox Tree, J. E. (1999). Listening in on monologues and dialogues. *Discourse Processes*, *27*(1), 35–53. https://doi.org/10.1080/01638539909545049

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/RSTB.2005.1622

Gervais, H., Belin, P., Boddaert, N., Leboyer, M., Coez, A., Sfaello, I., Barthélémy, C., Brunelle, F., Samson, Y., & Zilbovicius, M. (2004). Abnormal cortical voice processing in autism. *Nature Neuroscience 2004 7:8*, *7*(8), 801–802. https://doi.org/10.1038/nn1291

Golestani, N., Hervais-Adelman, A., Obleser, J., & Scott, S. K. (2013). Semantic versus perceptual interactions in neural processing of speech-in-noise. *NeuroImage*, *79*, 52–61. https://doi.org/10.1016/J.NEUROIMAGE.2013.04.049

Golestani, N., Rosen, S., & Scott, S. K. (2009). Native-language benefit for understanding speech-in-noise: The contribution of semantics*. *Bilingualism: Language and Cognition*, *12*(3), 385–392. https://doi.org/10.1017/S1366728909990150

Hagerman, B. (1984). Clinical Measurements of Speech Reception Threshold in Noise. *Scandinavian Audiology*, *13*(1), 57–63. https://doi.org/10.3109/01050398409076258

Iamshchinina, P., Karapetian, A., Kaiser, D., & Cichy, R. M. (2022). Resolving the time course of visual and auditory object categorization. *Https://Doi.Org/10.1152/Jn.00515.2021*. https://doi.org/10.1152/JN.00515.2021

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, 201714471. https://doi.org/10.1073/pnas.1714471114

Issard, C., Tsuji, S., & Cristia, A. (2023). *Infants' preference for speech is stable across the first year of life: Meta-analytic evidence*. https://doi.org/10.31222/OSF.IO/2BPMJ

Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., & Banerjee, S. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *116*(4), 2395. https://doi.org/10.1121/1.1784440

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Landsiedel, J., & Koldewyn, K. (2023). Auditory dyadic interactions through the 'eye' of the social brain: How visual is the posterior STS interaction region? *BioRxiv*, 2023.03.13.532398. https://doi.org/10.1101/2023.03.13.532398

Lee, J. H., & Humes, L. E. (2012). Effect of fundamental-frequency and sentence-onset differences on speech-identification performance of young and older adults in a competing-talker background. *The Journal of the Acoustical Society of America*, *132*(3), 1700. https://doi.org/10.1121/1.4740482

Levy, D. A., Granot, R., & Bentin, S. (2003). Neural sensitivity to human voices: ERP evidence of task and attentional influences. *Psychophysiology*, *40*(2), 291–305. https://doi.org/10.1111/1469-8986.00031

Li, G., & Lutman, M. E. (2006). *Sparseness and speech perception in noise*. http://www.isca-speech.org/archive/interspeech_2006/i06_1466.html

Magnus, B. E., Willoughby, M. T., Blair, C. B., & Kuhn, L. J. (2017). *Integrating Item Accuracy and Reaction Time to Improve the Measurement of Inhibitory Control Abilities in Early Childhood*.

Mattys, S. L., Brooks, J., & Cooke, M. (2009). Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive Psychology*, *59*(3), 203–243. https://doi.org/10.1016/J.COGPSYCH.2009.04.001

McArdle, R. A., Wilson, R. H., & Burks, C. A. (2005). Speech recognition in multitalker babble using digits, words, and sentences. *Journal of the American Academy of Audiology*, *16*(9), 726–739. https://doi.org/10.3766/JAAA.16.9.9/BIB

Miller, G. A. (1947). The masking of speech. *Psychological Bulletin*, *44*(2), 105–129. https://doi.org/10.1037/H0055960

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, *41*(5), 329–335. https://doi.org/10.1037/H0062491

Mulder, M. J., & Van Maanen, L. (2013). Are Accuracy and Reaction Time Affected via Different Processes? *PLOS ONE*, *8*(11), e80222. https://doi.org/10.1371/JOURNAL.PONE.0080222

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, *95*(2), 1085–1099. https://doi.org/10.1121/1.408469

Nordvik, O., Laugen Heggdal, P. O., Brännström, J., Vassbotn, F., Aarstad, A. K., & Aarstad, H. J. (2018). Generic quality of life in persons with hearing loss: A systematic literature review. *BMC Ear, Nose and Throat Disorders*, *18*(1), 1–13. https://doi.org/10.1186/S12901-018-0051-6/TABLES/2

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11*(12), 520–527. https://doi.org/10.1016/j.tics.2007.09.009

Olson, H. A., Chen, E. M.-Y., Lydic, K. O., & Saxe, R. R. (2023). Left-Hemisphere Cortical Language Regions Respond Equally to Observed Dialogue and Monologue. *Neurobiology of Language*, *4*(4), 575–610. https://doi.org/10.1162/NOL_A_00123

O'Neill, E. R., Parke, M. N., Kreft, H. A., & Oxenham, A. J. (2021). Role of semantic context and talker variability in speech perception of cochlear-implant users and normal-hearing listeners. *The Journal of the Acoustical Society of America*, *149*(2), 1224–1239. https://doi.org/10.1121/10.0003532

Papeo, L., & Abassi, E. (2019). Seeing social events: The visual specialization for dyadic human–human interactions. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(7), 877–888. https://doi.org/10.1037/xhp0000646

Papeo, L., Stein, T., & Soto-Faraco, S. (2017). The Two-Body Inversion Effect. *Psychological Science*, *28*(3), 369–379. https://doi.org/10.1177/0956797616685769

Parthasarathy, A., Hancock, K. E., Bennett, K., Degruttola, V., & Polley, D. B. (2020). Bottom-up and top-down neural signatures of disordered multi-talker speech perception in adults with normal hearing. *ELife*, *9*. https://doi.org/10.7554/ELIFE.51419

Peirce, 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*(1), 195–203. https://doi.org/10.3758/S13428-018-01193-Y/FIGURES/3

Pesquita, A., Rennels, J. L., Bernardet, U., Richards, B. E., Jensen, O., & Shapiro, K. (2022). Isolating Action Prediction from Action Integration in the Perception of Social Interactions. *Brain Sciences 2022, Vol. 12, Page 432*, *12*(4), 432. https://doi.org/10.3390/BRAINSCI12040432

Pickering, M. J., & Garrod, S. (2021). Understanding dialogue: Language use and social interaction. *Cambridge University Press*.

Plomp, R., & Mimpen, A. M. (1979). Improving the reliability of testing the speech reception threshold for sentences. *International Journal of Audiology*, *18*(1), 43–52. https://doi.org/10.3109/00206097909072618

Prinzmetal, W., McCool, C., & Park, S. (2005). Attention: Reaction time and accuracy reveal different mechanisms. *Journal of Experimental Psychology: General*, *134*(1), 73–92. https://doi.org/10.1037/0096-3445.134.1.73

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience 1999 2:1*, *2*(1), 79–87. https://doi.org/10.1038/4580

Roland, D., Yun, H., Koenig, J. P., & Mauner, G. (2012). Semantic similarity, predictability, and models of sentence processing. *Cognition*, *122*(3), 267–279. https://doi.org/10.1016/J.COGNITION.2011.11.011

Schelinski, S., Borowiak, K., & von Kriegstein, K. (2016). Temporal voice areas exist in autism spectrum disorder but are dysfunctional for voice identity recognition. *Social Cognitive and Affective Neuroscience*, *11*(11), 1812–1822. https://doi.org/10.1093/SCAN/NSW089

Simpson, S. A., & Cooke, M. (2005). Consonant identification in N-talker babble is a nonmonotonic function of N. *The Journal of the Acoustical Society of America*, *118*(5), 2775–2778. https://doi.org/10.1121/1.2062650

Skuk, V. G., Palermo, R., Broemer, L., & Schweinberger, S. R. (2019). Autistic Traits are Linked to Individual Differences in Familiar Voice Identification. *Journal of Autism and Developmental Disorders*, *49*(7), 2747–2767. https://doi.org/10.1007/S10803-017-3039-Y/FIGURES/8

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive Top-Down Integration of Prior Knowledge during Speech Perception. *Journal of Neuroscience*, *32*(25), 8443–8453. https://doi.org/10.1523/JNEUROSCI.5069-11.2012

Soleymani, R., Selesnick, I. W., & Landsberger, D. M. (2019). ALTIS: A new algorithm for adaptive long-term SNR estimation in multi-talker babble. *Computer Speech and Language*, *58*, 231–246. https://doi.org/10.1016/j.csl.2019.05.001

Stewart, M. E., & Ota, M. (2008). Lexical effects on speech perception in individuals with “autistic” traits. *Cognition*, *109*(1), 157–162. https://doi.org/10.1016/J.COGNITION.2008.07.010

Strauß, A., Kotz, S. A., & Obleser, J. (2013). Narrowed Expectancies under Degraded Speech: Revisiting the N400. *Journal of Cognitive Neuroscience*, *25*(8), 1383–1395. https://doi.org/10.1162/JOCN_A_00389

Strom, K. E. (2006). TheHR2006 dispensing survey. *Hearing Review*, *13*, 16–39.

Tolins, J., Zeamer, C., & Fox Tree, J. E. (2018). Overhearing Dialogues and Monologues: How Does Entrainment Lead to More Comprehensible Referring Expressions? *Discourse Processes*, *55*(7), 545–565. https://doi.org/10.1080/0163853X.2017.1279516

van Boxtel, J. J. A., Peng, Y., Su, J., & Lu, H. (2017). Individual differences in high-level biological motion tasks correlate with autistic traits. *Vision Research*, *141*, 136–144. https://doi.org/10.1016/j.visres.2016.11.005

van Ede, F., de Lange, F. P., & Maris, E. (2012). Attentional Cues Affect Accuracy and Reaction Time via Different Cognitive and Neural Processes. *Journal of Neuroscience*, *32*(30), 10408–10412. https://doi.org/10.1523/JNEUROSCI.1337-12.2012

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

Vestner, T., Gray, K. L. H., & Cook, R. (2021). Visual search for facing and non-facing people: The effect of actor inversion. *Cognition*, *208*, 104550. https://doi.org/10.1016/J.COGNITION.2020.104550

Vestner, T., Tipper, S. P., Hartley, T., Over, H., & Rueschemeyer, S.-A. A. (2019). Bound together: Social binding leads to faster processing, spatial distortion, and enhanced memory of interacting partners. *Journal of Experimental Psychology: General*, *148*(7), 1251–1268. https://doi.org/10.1037/xge0000545

Walbrin, J., & Koldewyn, K. (2019). Dyadic interaction processing in the posterior temporal cortex. *NeuroImage*, *198*(January), 296–302. https://doi.org/10.1016/j.neuroimage.2019.05.027

Wilcox, R. R. (2016). Comparing dependent robust correlations. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 215–224. https://doi.org/10.1111/BMSP.12069

Willoughby, M. T., Camerota, M., King, K. M., Nduku, T., & Piper, B. (2023). Leveraging item-level accuracy and reaction time to address ceiling effects in the measurement of inhibitory control in preschool-aged children. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/FPSYG.2023.861441/FULL

Wilson, R. H. (2003). Development of a speech-in-multitalker-babble paradigm to assess word-recognition performance. *Journal of the American Academy of Audiology*, *14*(9), 453–470. https://doi.org/10.1055/S-0040-1715938/BIB

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, and Psychophysics*, *79*(7), 2064–2072. https://doi.org/10.3758/S13414-017-1361-2/FIGURES/6

Wyer, N. A., Martin, D., Pickup, T., & Neil Macrae, C. (2012). Individual Differences in (Non-Visual) Processing Style Predict the Face Inversion Effect. *Cognitive Science*, *36*(2), 373–384. https://doi.org/10.1111/j.1551-6709.2011.01224.x

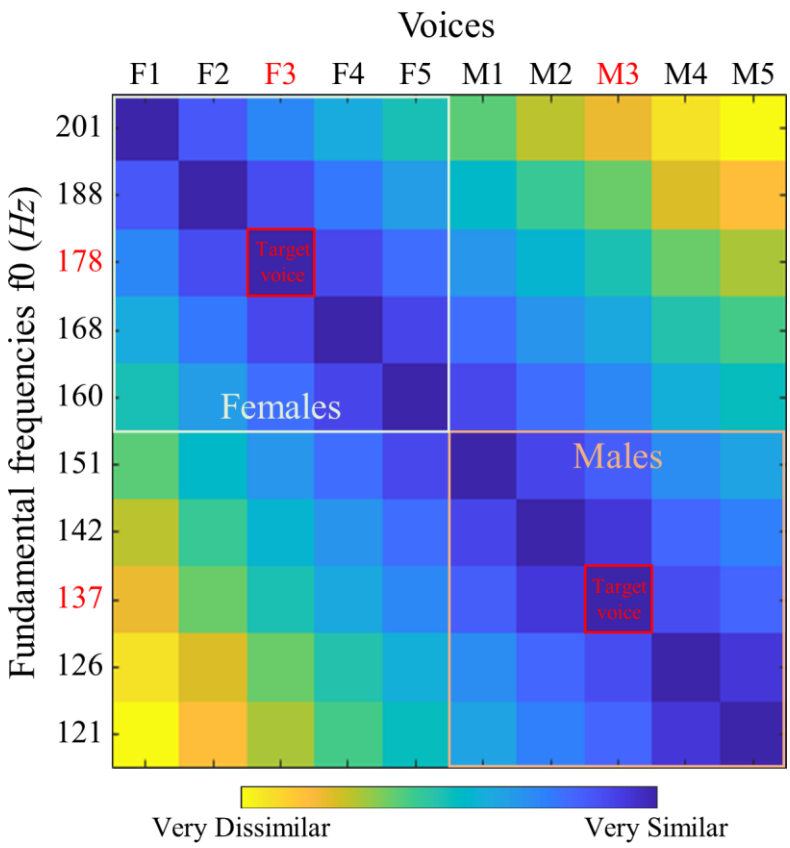
Zekveld, A. A., Rudner, M., Johnsrude, I. S., Festen, J. M., van Beek, J. H. M., & Rönnberg, J. (2011). The Influence of Semantically Related and Unrelated Text Cues on the Intelligibility of Sentences in Noise. *Ear and Hearing*, *32*(6), e16–e25. https://doi.org/10.1097/AUD.0B013E318228036A

Zekveld, A. A., Rudner, M., Johnsrude, I. S., Heslenfeld, D. J., & Rönnberg, J. (2012). Behavioral and fMRI evidence that cognitive ability modulates the effect of semantic context on speech intelligibility. *Brain and Language*, *122*(2), 103–113. https://doi.org/10.1016/J.BANDL.2012.05.006

Zekveld, A. A., Rudner, M., Johnsrude, I. S., Rönnberg, J., & R€, J. (2013). The effects of working memory capacity and semantic cues on the intelligibility of speech in noise. *The Journal of the Acoustical Society of America*, *134*(3), 2225–2234. https://doi.org/10.1121/1.4817926

Zhao, S., Brown, C. A., Holt, L. L., & Dick, F. (2022). Robust and Efficient Online Auditory Psychophysics. *Trends in Hearing*, *26*. https://doi.org/10.1177/23312165221118792/

Supplementary Materials



Supplementary Figure 1. Dissimilarity matrix for Euclidean distances between fundamental frequencies of the voices. Captions in red represent the target voices for female (F3) and male (M3). Captions in black represent the concurrent voices of females (F1, F2, F4, F5) and males (M1, M2, M4, M5) that are used in the multi-talker babbling noise.