Influence of social and semantic contexts on phonetic encoding in naturalistic conversations

Etienne Abassi[1] & Robert J. Zatorre[1]

[1]*Montreal Neurological Institute – McGill University,*

[1]*Centre for Research in Brain, Language and Music (CRBLM)*

[1]*International Laboratory for Brain, Music and Sound Research (BRAMS)*

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 significant part of life, and understanding others' conversations is key to navigating our social world. While the role of semantics in speech comprehension is well-established at the word or sentence level, its influence on larger conversational scales, alongside social context, is less understood. This study examined how semantic and social contexts modulate phonetic encoding during natural conversations using a speech-in-noise paradigm. Participants listened to AI-generated dialogues (two speakers) or monologues (one speaker) in intact or sentence-scrambled order. Each trial contained five sentences, with the fifth embedded in multi-talker babble noise. The same sentence was then repeated without noise, with one word either altered or unchanged. Healthy adults identified whether the sentence matched the in-noise version. Across several online experiments (N=195), both social and semantic contexts showed influences on speech-in-noise processing with improved performance for dialogues over monologues and for intact over sentence-scrambled conversations. These results suggest that both semantic and social factors shape speech comprehension, emphasizing their role in auditory cognition. This raises important questions about predictive and other mechanisms involved in processing complex, multi-sentence conversations, further underscoring the critical role of social interaction in communication.

KEYWORDS

Phonetic encoding, Conversations, Context, Social, Semantic

I. INTRODUCTION

During our lifetime, we humans have to manifest complex abilities to navigate our social world (Adolphs, 2001, 2009). Consequently, a significant portion of our existence is devoted to engaging in social interactions with others. Not only do we interact in the first person, but listening to others' interactions also serves critical functions in understanding our social world. Hence, successful understanding of others relies on the seamless integration of sensory input and cognitive mechanisms, enabling us to understand speech across diverse and often challenging auditory environments. Central to this process is phonetic encoding, the mapping of acoustic signals onto linguistic representations, which underpins speech comprehension (Yi et al., 2019; Lahiri & Reetz, 2010; Clements 1985; Jakobson et al., 1951; Sapir, 1925). While much of the existing research has examined how speech is processed in controlled environments using single sentences or words, real-world communication involves multiple layers of complexity, including listening contexts that influence how phonetic encoding unfolds (Kuperberg & Jaeger, 2016; Sohoglu et al., 2012). Understanding the mechanisms by which these contexts shape speech perception is critical for advancing theories of auditory cognition and improving interventions for individuals with hearing loss.

Semantic context, or the coherence and meaning embedded in speech, has long been recognized as a critical factor in speech processing, and plays a pivotal role in phonetic encoding. Coherent sentences allow listeners to predict upcoming words and phonetic structures, reducing cognitive load and improving processing efficiency (Sohoglu et al., 2012; Zekveld et al., 2011, 2012, 2013). For example, sentences with meaningful content enable listeners to disambiguate phonetic details, particularly in acoustically challenging environments such as noisy or multi-talker settings (Kuperberg & Jaeger, 2016; O'Neil et al., 2021). These predictive mechanisms are

often conceptualized within the framework of predictive coding, which posits that the brain generates expectations based on prior input to minimize the mismatch between sensory input and cognitive predictions (Clark, 2013; Friston, 2005; Rao & Ballard, 1999). Semantic congruence facilitates phonetic encoding by reinforcing expectations and enhancing signal clarity, while semantic incongruence or nonsense sentences disrupt these processes and impose greater cognitive demands (Golestani et al., 2009, 2013; Zekveld et al., 2011).

Research on semantic context has typically focused on isolated words or sentences. For instance, studies have shown that listeners process phonemes more accurately when words are embedded in semantically predictable sentences (Roland et al., 2012; Zekveld et al., 2013) and that speech patterns at different timescales are processed using different mechanisms and/or brain regions (DeWitt & Rauschecker, 2012; Farbood et al., 2015). Additionally, experiments involving word priming or manipulation of word probabilities within sentences have demonstrated that semantic expectations also modulate auditory processing at early sensory stages (Broderick et al., 2019). These findings suggest that the integration of semantic cues operates at multiple levels, from immediate phonetic recognition to broader linguistic interpretation. Despite this robust evidence, relatively little is known about how semantic context functions at the scale of whole conversations, where coherence emerges over longer timescales. Conversations, unlike isolated sentences, involve complex semantic interdependencies that unfold dynamically, allowing for greater predictive stability. Studies examining connected speech and multi-sentence passages, such as those using the Connected Speech Test (Cox et al., 1987) or naturalistic corpora like the LUCID database (Baker & Hazan, 2010, 2011), suggest that broader semantic contexts enhance speech processing. However, these studies often focus on general intelligibility rather than phonetic encoding, leaving a gap in understanding how semantic coherence shapes the underlying processes

of speech perception. As such, a primary aim of the current study is to address this gap by examining how semantic context influences phonetic encoding in multi-sentence conversations, providing insights into the broader cognitive and linguistic mechanisms that support speech processing.

While semantic context focuses on the linguistic coherence of speech, social context, on the other hand, encompasses the interactional dynamics and relational cues present in communication. As such, it may represent another critical factor influencing phonetic encoding. Dynamic exchanges between speakers, such as turn-taking and conversational flow, may engage cognitive systems that are tuned to socially relevant stimuli while providing listeners with additional scaffolding for auditory processing. For instance, dialogues involving alternating speakers provide structural and temporal cues that enhance speech comprehension compared to monologues (Fox Tree, 1999). These interactional dynamics may facilitate phonetic encoding by directing attention to relevant acoustic signals and reinforcing cognitive predictions about speaker behavior (Levy et al., 2003). Social context is further enriched by factors such as speaker identity, familiarity, and social relationships. Familiar voices are processed more efficiently than unfamiliar ones (Johnsrude et al., 2013; Domingo et al., 2020), as listeners can leverage prior knowledge to disambiguate phonetic details and interpret speech more effectively (Belin et al., 2000; McGowan, 2015). Similarly, shared social knowledge and relational dynamics can modulate auditory processing by providing additional contextual scaffolding. For example, listeners interpret speech more accurately when they share a common cultural or linguistic background with the speaker (Lev-Ari & Keysar, 2010; Rubin, 1992). These effects suggest that social interaction engages specialized neural and cognitive systems that enhance phonetic encoding and speech comprehension. Dialogues, as a prototypical form of social interaction, offer particularly rich

opportunities for examining the influence of social context. Unlike monologues, dialogues involve interactive turn-taking, which creates temporal predictability and facilitates attentional engagement. Previous studies have shown that listeners perceive dialogues as more intelligible than monologues, even when the content is identical, suggesting that interactional dynamics enhance cognitive processing (Branigan et al., 2011; Pickering & Garrod, 2021; Tolins et al., 2018). However, the mechanisms underlying this advantage remain poorly understood.

Going further, the influence of context in sensory processing 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 the neural representation of a single body within a dyad is sharpened when the two bodies are represented face-to-face (as if interacting) as compared to back-to-back (Abassi and Papeo, 2020; Bellot et al., 2021). This shows that a given social context can have a significant influence on the perception of a visual stimulus and its neural processing. Still, 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. Thus, building on the literature of both social audition and vision, a major aim of this study is isolating the effects of social interaction on phonetic encoding, controlling for semantic content to focus on the contribution of interactional cues such as speaker alternation and conversational flow.

In addition to interactional dynamics, social context can also modulate speech processing through social evaluation. Indeed, listeners' perceptions of intelligibility are influenced by social factors such as stereotypes, biases and attitudes toward the speaker (Babel & Russell, 2015; Lev-Ari & Keysar, 2010; Niedzielski, 1999; Rubin, 1992). For example, speech produced by non-native speakers or individuals from different social groups may be perceived as less intelligible due to listeners' preconceived notions or social judgments, even when the acoustic properties are similar (Babel & Russell, 2015; McGowan, 2015). While not the primary focus of this study, these findings highlight the broader role of social cognition in shaping auditory perception and underscore the importance of considering social influences when investigating phonetic encoding. They also highlight the interplay between social cognition and speech perception, suggesting that social evaluations can modulate auditory processing at various levels (Drager & Kirtley, 2016; Niedzielski, 1999). Hence, here we provide a new approach of this question by focusing on how social interaction dynamics, rather than social evaluations or biases, influence phonetic encoding.

On a last point, individual differences in sensitivity to social contexts may influence phonetic encoding, particularly through variations in social traits. Indeed, the Autism Spectrum Quotient (AQ), a widely used measure of autism-related traits in both neurodiverse and neurotypical populations (Baron-Cohen et al., 2001), has been consistently linked to reduced sensitivity to social stimuli. More specifically, neurotypical individuals with higher AQ scores demonstrate diminished sensitivity in visual tasks involving social stimuli, such as processing dyadic body interactions (Abassi & Papeo, 2022), faces (Wyer et al., 2012), biological motion (van Boxtel et al., 2017), and animals (Yang et al., 2024). These findings suggest a general link between individual differences in social-related traits and integration of social cues in the visual modality, which may extend to auditory perception in contexts emphasizing social interaction dynamics. In

parallel, higher AQ scores have been associated with enhanced attention to fine phonetic details (Ota et al., 2015; Yu, 2010; Steward & Ota, 2008), suggesting a potential advantage in auditory tasks focusing on precise acoustic distinctions. However, this strength may not fully compensate for the reduced sensitivity to social dynamics in tasks requiring the integration of social cues for phonetic processing. This dual influence highlights how individuals with higher AQ scores might excel in isolated phonetic tasks but face challenges in contexts where social dynamics are critical for guiding perceptual processes. By including AQ scores in the present study, we aim to investigate how sensitivity to social interaction dynamics may modulate the effects of semantic and social contexts on phonetic encoding.

Overall, in the current study we investigate the influence of semantic and social contexts on phonetic encoding. Specifically, we hypothesize that both semantic coherence and dynamic social interactions would enhance the encoding of phonetic information. In addition, while we predict an effect for each of these factors, we do not know if they are independent or interacting, that is another question we address in this study. To test these hypotheses, we manipulated the semantic context (intact vs. sentence-scrambled conversations) and social context (dialogues vs. monologues) in which a sentence was heard, while using a speech-in-noise paradigm as a tool to measure perceptual sensitivity to this sentence. Speech-in-noise paradigms provide an effective framework for studying phonetic encoding under controlled conditions that approximate real-world listening challenges. Indeed, these paradigms introduce competing auditory stimuli that require listeners to rely on higher-order cognitive mechanisms, such as context-based predictions in our case, to process speech effectively (Zekveld et al., 2013). In the current study, by combining speech-in-noise tasks with manipulations of semantic coherence and social interactions dynamics, we aim to isolate the mechanisms by which these contexts enhance phonetic encoding. However,

unlike previous research, which has often focused on isolated words or sentences, our study examines how these factors operate at the level of larger-scale naturalistic conversations, offering a more comprehensive understanding of real-world speech perception. The implications of this research extend beyond basic auditory science, as individuals with hearing loss or social-cognitive impairments often struggle to process speech in complex environments where both semantic and social cues play a critical role (Ciorba et al., 2012; Nordvik et al., 2018). By elucidating the mechanisms underlying phonetic encoding in these contexts, this study lays the groundwork for developing more ecologically valid assessment tools and assistive technologies, such as hearing aids that leverage conversational dynamics and contextual predictions to enhance speech processing.

II. METHODS

A. Participants. A total of 295 participants took part in our on-line study (Exp 1: 80 participants, 29 females, 3 sex unreported, mean age 29.8 years +/- 6.2 SD; Exp 2: 215 participants, 85 females, 2 sex unreported, mean age 31.6 years +/- 5.6 SD). Overall, the minimal sample size for each experiment was based on previous literature using analogous word-recognition tasks in multi-talker babble noise. More specifically, Killion et al. (2004) found significant results for the QuickSIN test using 16 participants and Wilson (2003) used 24 participants for the WIN test, while in our current study we had a minimum of 40 participants for each experimental subset (see below). All participants read and signed an informed consent form prior to participating in the study and were remunerated 6£ for ~40 minutes of their time, paid through the on-line platform (prolific.com). They were all American English first-language speakers with normal hearing (determined by self-reported online), aged between 18 and 35 years.

B. Stimuli. The entire set of stimuli can be found online at www.zlab.mcgill.ca/stimuli/ ('Stimuli_from_Abassi_2025').

Conversational stimuli. The main stimulus 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. Each conversation was carefully constructed to ensure that dialogues and monologues had the same sentence structures and themes, isolating the effect of social context. Indeed, dialogues featured alternating male and female voices, providing turn-taking cues and interactive dynamics that were absent in monologues, which were all produced by the same voice. This design allowed us to investigate how the presence of two interacting voices, versus a single continuous voice, influences the listener's processing of 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 5 sentences each between two persons with a maximum of 15 words per sentence"*. We did not use the exact conversation contents as generated by ChatGPT, but rather made modifications such as word, syntax, and length changes to enhance the naturalistic aspect of each item. For each five-sentence conversation, we created a sixth sentence by hand, that was same as the fifth sentence, but with only one word change that could appear around the beginning, the middle or the end of the sentence. The choice of the positions for the word changes was made so it was sufficiently balanced across the three possibilities (begin, middle or end). The changed word 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, hence no semantic anomaly was introduced; see Fig 1), and had same number of syllables as the target word. Thus, the manipulations were exclusively related to the *context* within which the target sentence was presented, and not to the structure or content of the target sentence itself.

Prior to converting the conversations to natural-sounding speech, we added prosody tags in each sentence to make them 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>* ». The complete list of sentences with their prosody tags is available in the stimuli dataset package. 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 root mean squared (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 (i.e., across distinct topics), 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). Between every sentence, we added a silent gap randomized between 400 and 600 ms, except between the fifth and the sixth sentences,

where a silent gap of 1 s. was added. On average, each full conversation had a total duration of ~18 s.

Figure 1. Example of a conversation 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 green standard text alternating with brown italic text represent the alternation 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. The entire set of stimuli can be found online at *www.zlab.mcgill.ca/stimuli/* ('Stimuli_from_Abassi_2025').

One-sentence stimuli. In addition to dialogues and monologues, we created a set of 50 one-sentence stimuli that were different than the sentences used in the conversations and were used as a baseline condition to control for inter-individual variability in performance in processing speech-in-noise. For each of these sentences, we created a second sentence by hand, that was same as the first sentence, but with only one word change that could appear around the beginning, the middle

or the end of the sentence. Similarly, as for the main experiment (conversations), 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. We then converted these sentences into natural-sounding speech for both female and male voices, using the same method as for the conversations. For example, one 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 15 words per sentence".* 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), for an average duration of ~5 s. each.

C. Multi-talker babble noise. In a second step, we embedded the fifth sentence of each conversation, as well as all the one-sentence baseline stimuli into multi-talker babble noise. We used multi-talker babble 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).

Noise features. The features of the multi-talker babble noise was set 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 they would not overlap with the target voices, as having too close f0 between target and competing voices may interfere too much 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 supplementary Figure 1), so the range would be large enough to cover a representative range of f0 in the general population (see Cristina Oliveira et al., 2021).

Noise distractor sentences generation. To generate the content of the noise distractor sentences, we 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 correct grammar (Van Engen & Bradlow, 2007). An example of the ChatGPT prompts used is *"Write X semantically anomalous but syntactically correct 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. Using Matlab, 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.

Embedding of target sentences in noise. After generating the 24 audio distractor 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 complete noise tracks was faded in. In total, we produced 120 multi-talker noise stimuli. In a final step, we mixed the fifth target 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 target sentences with the addition of 100ms before and after the sentences during which there was only noise. Here, the level of the multi-talker masking noise was held constant, while the amplitude of the target stimuli was decreased by SNRs of -1, -4, and −6 dB, where -4 and -6 dB are known to produce a speech reception threshold (see Plomp & Mimpen, 1979) that is in the range of 40% to 60% depending on the studies and methods (Brungart et al., 2001; Dingemanse & Goedegebure, 2020; Festen & Plomp, 1990; Hagerman, 1984; McArdle et al., 2005; Miller et al., 1951). These values were also tested in prior in-the-lab pilot testing (unpublished data) to estimate a range of SNRs that would adequately capture the skill level of young healthy participants, while limiting floor or ceiling effects as well as subject frustration (Coffey et al., 2019).

D. Procedure. We used a speech-in-noise paradigm to study how semantic and social contexts interact with phonetic encoding of a sentence. Importantly, we used a task that 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). This manipulation allowed us to focus on the phonetic encoding of a word rather than its underlying meaning, moving beyond traditional measures of intelligibility by focusing on the deeper cognitive processes that enable robust speech perception.

Experiments rationales. We conducted two main experiments. Experiment 1 was a preliminary study designed to select a correct SNR (-4 dB or -6 dB) to be used in the subsequent experiments and was composed of two participant groups: one with dialogues only (Exp. 1a) and one with monologues only (Exp. 1b). Experiment 2 addressed the effects of semantic and social contexts while controlling for effects of different experimental conditions/orders (Carvalho & Goldstone, 2014a, 2014b; Francis & Ciocca, 2003), and was composed of 3 participant groups: Experiment 2a was composed of separate blocks of dialogues-only trials and blocks of monologues-only trials, each featuring both intact and sentence-scrambled conversations within each block; Experiment 2b was composed of separate blocks of intact conversations only and blocks of sentence-scrambled conversations only, each featuring both dialogues and monologues within each block; and Experiment 2c was composed of blocks featuring all four conditions within each block. All experiments (for which details are given below) 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 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 avoid having any object around them that may disturb their attention, such as a smartphone or a TV on.

Check of optimal on-line experimental conditions. Prior to beginning an experiment, 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 babble noise alone without target sentences and they were instructed to

adjust their audio volume to 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 -6 dB 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. If participants failed 5 out of 6 trials, they were proposed to redo the headphone check a second time. If they failed the headphone check a second time, they were not allowed to carry on with the experiment and the study automatically stopped for them.

Experimental procedure. 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 are 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. Each experiment consisted of two main tasks: 1) a baseline task, in which only one-sentence stimulus was presented in noise, that was used to control the level of understanding of sentences-in-noise for each

participant 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. 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 was to determine, as quickly as possible, whether the repeated sentence was exactly the same or different than the one presented in noise. 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 stimulus, then by a visual reminder of the response buttons until the participants pressed one of the two buttons. 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 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).

All responses (Accuracy, reaction time [RTs] and AQ questionnaire) were collected with Pavlovia and processed with Matlab. Measure of reaction times began directly after the last sentence was heard until the participant pressed one of the two response buttons.

Experiment 1. Experiment 1 was conducted to choose a correct SNR for measuring our effects of interest in the subsequent experiments and included two groups, with one group presented with dialogues only (Exp. 1a; 40 participants) and one group presented with monologues only (Exp. 1b;

40 participants). Here, we expected a correct SNR to not produce ceiling of floor effects with accuracy scores and to produce our effect of interest (i.e. better performance for intact over sentence-scrambled conversations using d' and/or RTs). For each group, half or the unique stimuli were presented with an SNR of -6dB, while the other half was presented with and SNR of -4dB, and stimuli were randomly assigned in one or the other half for each participant. 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. 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.

Experiment 2a. Experiment 2a included 74 participants and was composed of blocks of dialogue only or monologues only, while a same block featured both intact and scrambled conversations. For each participant, half of the unique stimuli was presented in the dialogue condition and repeated across the intact and scrambled conditions while the other half was presented in the monologue condition and repeated across the intact and scrambled conditions, and stimuli were randomly assigned in one or the other half for each participant using 6 predefined lists. Experiment 2a used an SNR of -4 dB and was conducted as follows. At first a training task featured 6 baseline trials, followed by a testing task of 44 baseline trials. 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.

Experiment 2b. Experiment 2b included 66 participants and was composed of blocks of intact conversations only or sentence-scrambled conversations only, while a same block featured both dialogues and monologues. For each participant, half of the unique stimuli was presented in the intact condition and repeated across dialogues and monologues while the other half was presented in the scrambled condition and repeated across dialogues and monologues, and stimuli were randomly assigned in one or the other half for each participant using 6 predefined lists. Experiment 2b used an SNR of -4 dB and was conducted as follows. At first a training task featured 6 baseline trials, followed by a testing task of 44 baseline trials. 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.

Experiment 2c. Experiment 2c included 75 participants and was composed of blocks featuring all conditions randomly intermixed (i.e., both dialogues and monologues in both intact and sentence-scrambled orders, not blocked into conditions). For each participant, half of the unique stimuli were presented in the dialogue condition and repeated across the intact and scrambled conditions while the other half were presented in the monologue condition and repeated across the intact and scrambled conditions, and stimuli were randomly assigned in one or the other half for each participant using 6 predefined lists. Experiment 2c 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. 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.

Autism Quotient (AQ) task. For experiment 2, at the end of each acquisition 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 ASD (Autistic Spectrum Disorders), 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". Calculation of the AQ quotient was performed following the instructions provided by Baron-Cohen et al. (2001). Higher scores indicate greater ASD traits while lower scores indicate fewer ASD traits.

E. Data preprocessing. For all experiments, the following data preprocessing steps were applied: 1) For each participant, trials for which RTs were >3 SD above the participant mean were discarded from further analyses; 2) For each participant, trials in which the response was inaccurate were discarded for RTs only; 3) Participants were excluded from the study if their mean RTs averaged across all conversation conditions was >3 SD above the subgroup mean (Participants exclusions per experiment: *Exp. 1a:* 1/40; *Exp. 1b:* 1/40; *Exp. 2a:* 2/74; *Exp. 2b:* 1/66; *Exp. 2c:* 1/75; *Total of participants exclusions:* 6/295; ~2.03 %). After exclusions, statistical analyses included 39 participants for experiment 1a, 39 participants for experiment 1b and 211 participants for experiment 2.

F. Analyses. For all experiments, d-prime scores (*d'*) and mean RTs were used to compare the effect of intact vs. scrambled conversation across dialogues and monologues. *d'* values were computed with the *psyphy* package (Knoblauch, 2023) run through R studio (R Core Team, 2023), using a differencing approach (Hautus et al., 2021). For Exp. 1, we additionally computed accuracy

to compare potential ceiling or floor effects as a function of the SNR. Then, for exp. 1, separately for *d'* and RTs we ran two linear mixed effect models (one for Exp. 1a with dialogues only and one for Exp. 2b with monologues only) with 2 factors [Semantic context (Intact / Scrambled) x SNR (-4 dB / -6 dB)], baseline as covariate and subjects as random factor. For experiment 2, separately for d' and RTs we ran one linear mixed effect model (LMM) with 3 factors [Experimental design (a / b / c) x Semantic context (Intact / Scrambled) x Social context (Dialogue / Monologue)], baseline as covariate, and subjects as random factor. For each of these models, baseline performance was included as a covariate to take into account individual variability in performing the speech-in-noise task prior to manipulating the context. In addition, we also ran the same LMM analysis for experiment 2 without including the baseline as a covariate, and we found that the model fit was better with the baseline than without the baseline (supplementary table 1), although with very similar results (supplementary table 2). Finally, when appropriate, post-hoc analyses were performed using pairwise *t* tests (two-tailed).

In addition, to study whether we could find a relationship between individual differences in performance in processing conversations and each individual's autistic traits measured through the AQ, separately for dialogues and monologues we employed a one-tail partial Spearman correlation between the individual AQ scores and the individual difference scores for intact minus scrambled conversations while controlling for the potential confounding effect of baseline performance. Here, we used a one-tail test because we had a clear hypothesis regarding the direction of our expected effect, a negative correlation, that would show that the lower the AQ score, the higher the difference of performance between intact and scrambled conversations. This hypothesis was directly driven from the results of an analogous study in the visual modality (Abassi & Papeo, 2022), that showed a negative correlation between individual AQ score and a

measure of behavioral performances in visual processing of facing body dyads (as if interacting) as compared to non-facing body dyads, as well as other studies that showed similar relationship between AQ and visual sensitivity to other type of social stimuli (van Boxtel et al., 2017; Wyer et al., 2012; Yang et al., 2024). We then compared the difference of magnitudes of the Spearman *rhos* (Lenhard & Lenhard, 2014) between dialogues and monologues.

III. RESULTS

A. Experiment 1. Experiment 1 was conducted to choose a correct SNR for measuring our effects of interest in the subsequent experiments. To do so, based on prior in-the-lab pilots (unpublished data) we selected two SNRs of -6 and -4 dB. For the SNR of -6 dB we found a mean baseline accuracy of 61.9 % (+/- 1.9 SEM) for Exp. 1a and 64.5 % (+/- 1.8 SEM) for Exp. 1b, while for the SNR of -4 dB, we found a mean baseline accuracy of 73.2 % (+/- 1.8 SEM) for Exp. 1a and 72.3 % (+/- 1.8 SEM) for Exp. 1b. For the main task, for the SNR of -6 dB we found a mean average accuracy of 61.7 % (+/- 1.8 SEM) for Exp. 1a and 60.2 % (+/- 1.9 SEM) for Exp. 1b, while for the SNR of -4 dB we found a mean average accuracy of 78.1 % (+/- 1.8 SEM) for Exp. 1a and 77.1 % (+/- 1.5 SEM) for Exp. 1b.

Using d', for Exp. 1a (dialogues only), the 2 SNR x 2 Semantic context LMM (fig. 2a) showed a main effect of SNR and no main effect of semantic context. Critically, it did show a significant interaction between SNR and Semantic context (table 1a). This interaction arises from a larger difference for intact vs. scrambled conversations for a SNR of -4 dB as compared to a SNR of -6 dB. Post-hoc *t*-tests showed that this difference was significantly higher for intact than scrambled conversations for a SNR of -4 dB ($t(38) = 2.49$, $p = 0.017$), while there was no difference for a SNR of -6 dB ($t(38) = 0.35$, $p > 0.250$). Using RTs (fig. 2b), we didn't find any significant

effects (table 1b). For Exp. 1b (monologues), we didn't find any effects using *d'* (table 1c; fig. 2c) or RTs (table 1d; fig. 2d).

To summarize, we first found that an SNR of -6 dB was producing accuracies in the range of 60-65%, while an 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 be too close to chance level, while and SNR of -4 might be appropriate for our needs. In addition, only with a SNR of -4 dB we found our effect of interest using d' (i.e., a significantly better performance for intact vs. scrambled dialogues in Exp. 1a). For this reason, we used an SNR of -4 dB for all subsequent experiments.

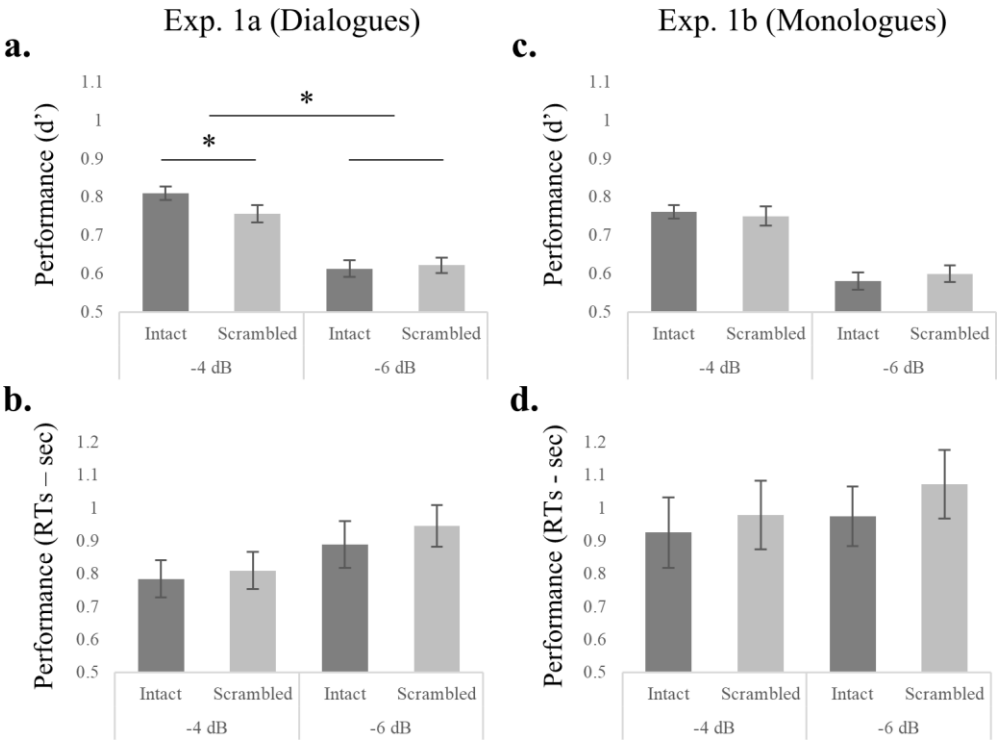


Figure 2. Behavioral data for experiment 1. Performances for intact and sentence-scrambled conversations using d-primes (d') and response times (RTs) with SNRs of -4 dB and -6 dB for Exp. 1a (Dialogues only: **a.**; **b.**) and Exp. 1b (Monologues only: **c.**; **d.**). Error bars represent standard errors of the mean. *$p \leq 0.05$.

a. Dialogues - d'

	F	df	df (res)	p
SNR	37.94	1	44.7	**<.001**
Semantic	1.53	1	38	0.223
Baseline	17.07	1	67.3	**<.001**
SNR x Semantic	4.18	1	38	**0.048**

c. Monologues - d'

	F	df	df (res)	p
SNR	58.76	1	36.7	**<.001**
Semantic	0.06	1	73	0.802
Baseline	5.83	1	73.9	**0.018**
SNR x Semantic	0.72	1	76	0.398

b. Dialogues - RTs

	F	df	df (res)	p
SNR	1.40	1	38.9	0.243
Semantic	1.62	1	75.6	0.207
Baseline	41.82	1	46.5	**<.001**
SNR x Semantic	0.23	1	76	0.635

d. Monologues - RTs

	F	df	df (res)	p
SNR	2.31	1	39.5	0.137
Semantic	2.68	1	38	0.110
Baseline	189.33	1	48.2	**<.001**
SNR x Semantic	0.34	1	38	0.564

Table 1. Results of the Linear Mixed Effect models with d' and RTs for dialogues (Exp. 1a: a.; b.) and monologues (Exp. 1b: c.; d.).

B. Experiment 2. Using *d'* (fig. 3a; Table 2a), the 3 Design x 2 Semantic context x 2 Social context LMM showed only a significant main effect of semantic context, characterized by higher d-primes for intact than sentence scrambled conversations, and no interactions. Post-hoc *t*-tests showed that the main effect of semantic context was qualified by a higher d' for intact vs. scrambled conversations for monologues only (dialogues: $t(210) = 1.02$, $p > 0.250$; monologues: $t(210) = 2.74$, $p = 0.007$).

Using RTs (fig. 3b; Table 2b) the same analysis showed a significant effect of social context, characterized by lower RTs for dialogues than monologues, as well as a significant effect of semantic context, characterized by lower RTs for intact than sentence-scrambled conversations. Post-hoc *t*-tests showed that the main effect of semantic context was qualified by lower RTs for intact as compared to sentence-scrambled conversations for both Dialogues and Monologues (Dialogues: $t(210) = 2.70$, $p = 0.007$; Monologues: $t(210) = 2.02$, $p = 0.045$). It also showed a trend

for an interaction between experimental design and social context, but no other interactions. Regarding the marginal interaction of experimental design and social context, post-hoc *t*-tests showed that the effect of social context was mainly driven by experiment 2c, in which all the experimental conditions were randomly presented within the same blocks as, when take separately, it was the only one with significantly lower RTs for dialogues as compared to monologues (Exp. 2a: $t(71) = 0.14$, $p > 0.250$; Exp. 2b: $t(64) = 0.82$, $p > 0.250$; Exp. 2c: $t(73) = 3.30$, $p = 0.001$).

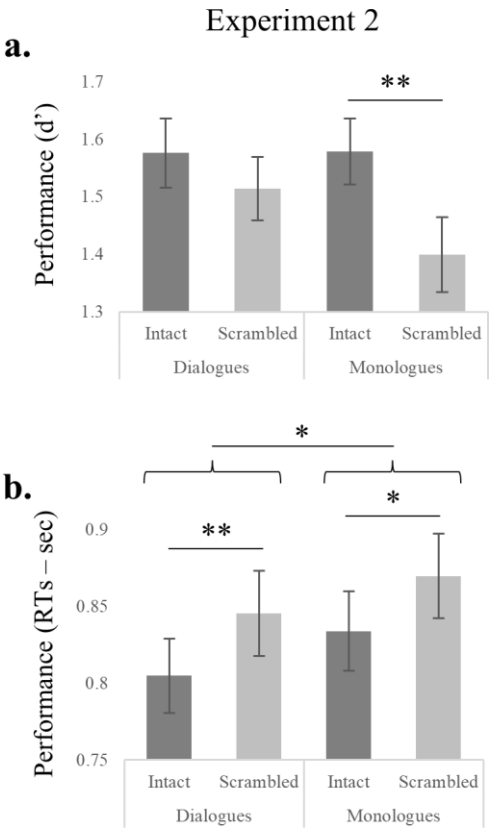


Figure 3. Behavioral data for experiment 2. Performances for intact and sentence-scrambled dialogues and monologues using d-primes (**a.**) and response times (**b.**). Error bars represent standard errors of the mean. *p ≤ 0.05; **p ≤ 0.01.

a. d'

	F	df	df (res)	p
Design	0.42	2	207	0.655
Social	1.60	1	384	0.207
Semantic	7.35	1	210	**0.007**
Baseline	75.54	1	207	**<.001**
Design x Social	0.04	2	384	0.956
Design x Semantic	0.40	2	210	0.670
Social x Semantic	1.75	1	416	0.186
Design x Social x Semantic	0.39	2	416	0.679

b. RTs

	F	df	df (res)	p
Design	1.33	2	207	0.266
Social	4.40	1	208	**0.037**
Semantic	9.55	1	208	**0.002**
Baseline	154.55	1	207	**<.001**
Design X Social	2.84	2	208	*0.061*
Design X Semantic	1.59	2	208	0.206
Social X Semantic	0.02	1	208	0.889
Design X Social X Semantic	0.75	2	208	0.472

Table 2. Results of the Linear Mixed Effect models with d' (a.) and RTs (b.) for Experiment 2.

C. Relationship between individual AQ scores and phonetic encoding. We then studied whether we could find a relationship between individual AQ scores and performances in processing conversations (performances for Intact minus Scrambled conversations) as a function of the social context (dialogues vs. monologues), using the data of experiment 2. We didn't find significant partial spearman correlations using the *d'* measure (Dialogues; *rho*(211) = -0.04; $p > 0.250$; Monologues; *rho*(211) = 0.07; $p = 0.161$) with no differences between dialogues and monologues ($z = -1.1$ $p = 0.136$). However, using RTs, only for dialogues and not for monologues we found that there was a trend for a negative correlation between AQ and processing of intact vs. scrambled conversations (Dialogues: rho(211) = -0.10, p = 0.067; Monologues: rho(211) = 0.02, p > 0.250). Furthermore, we also found a trend for a lower correlation for dialogues as compared to monologues ($z = -1.38$ $p = 0.085$).

IV. DISCUSSION

We investigated how semantic and social contexts modulate phonetic encoding during processing of naturalistic conversations using a novel speech-in-noise task. First, we found that semantic context significantly enhanced phonetic encoding. Specifically, participants performed better (lower response time and higher accuracy) when processing semantically coherent conversations compared to sentence scrambled ones. Second, social context also had a significant effect, with dialogues providing greater benefits (lower response time, but no effect on accuracy) to phonetic encoding than monologues. Finally, we found a trend for a negative correlation between individual differences in AQ traits and performances in processing intact vs. scrambled for dialogues only.

Our finding that semantic coherence within a full conversation enhances phonetic encoding aligns well with existing research on predictive coding theories of auditory perception. Predictive coding posits that the brain generates top-down expectations based on semantic context to minimize mismatches between sensory input and prior knowledge (Clark, 2013; Friston, 2005). In our study, participants performed better when processing semantically coherent conversations, demonstrating the benefits of predictive mechanisms in noisy environments where bottom-up auditory signals are degraded. These results are consistent with previous work showing that semantic congruence facilitates words and phonemes identification while reducing cognitive load (Golestani et al., 2009, 2013; Sohoglu et al., 2012; Zekveld et al., 2011, 2012, 2013). However, our study extends these findings into a new domain by showing effects of semantic context at the larger scale of naturalistic, multi-sentence conversations. Here, while prior research has established the role of semantic context within short linguistic units (Broderick et al., 2019; Roland et al., 2012), our results suggest that predictive mechanisms also operate over longer timescales,

integrating information across whole conversations. Indeed, in our paradigm, unlike most prior studies, the target sentence embedded in noise was not itself semantically anomalous; it was only anomalous or not with respect to the preceding conversational context.

While much of the prior research has framed these effects within the predictive coding framework, other complementary mechanisms may also explain the influence of semantic context on speech processing. First, semantic priming effects, whereby exposure to semantically related words activates shared neural networks, may also facilitate phonetic encoding by reinforcing linguistic expectations (Daltrozzo et al., 2011; Foss, 1982; Kutas & Federmeier, 2011). This process likely operates at multiple levels of representation, from lexical access to sentence-level coherence, allowing listeners to extract meaning even when speech is partially masked by noise. Secondly, semantic context may also interact with attentional mechanisms, prioritizing coherent information over incongruent or irrelevant input. Research on the cocktail party effect has shown that semantic predictability can help listeners segregate target speech from competing noise, even in multi-talker environments (Bronkhorst, 2015; Tóth et al., 2020). These findings suggest that semantic coherence not only aids phonetic encoding but also enhances the perceptual organization of auditory scenes, contributing to the broader goal of communication in complex environments. Interestingly, our study highlights the robustness of semantic effects across whole conversations, suggesting that listeners dynamically integrate information over longer timescales. This temporal dimension aligns well with research showing that discourse-level processing involves distinct neural mechanisms, including prefrontal and temporo-parietal networks, that support integration of meaning across sentences (Kutas & Federmeier, 2011; Menenti et al., 2012). Future research should explore how these long-range mechanisms interact with immediate predictive processes to shape phonetic encoding.

Our second and more novel finding pertains to social context effects, that showed better performances for dialogues over monologues. This result extends prior research emphasizing the role of interactional dynamics in speech processing. While the benefits of turn-taking and temporal predictability in dialogues are well-documented (Branigan et al., 2011; Fox Tree, 1999; Menenti et al., 2012; Pickering & Garrod, 2021; 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 can 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 vocal timbre of the speakers and their alternation 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, instead than emphasizing the content of the dialogue, we rather hypothesize that the presence of social cues, such as two speakers interacting with each other, triggers a sharpening of the acoustics of the sentences by the perceptual system, that is emphasizing some features of the auditory signal to facilitate comprehension, that would be higher for dialogues over monologues. This account suggests in turn that dialogues would provide additional cognitive engagement through social cues such as turn-taking and speaker interaction, which are absent in monologues. Hence, dialogues, due to their interactive nature, would create a richer social context that may engage listeners more deeply by eliciting greater social attention and enhancing speech

processing. This process could be mediated by neural systems specialized for processing socially relevant stimuli, such as the superior temporal sulcus (STS) and medial prefrontal cortex (Schilbach et al., 2006, 2008). 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., mimicking a social interaction), as compared to facing in an opposite direction (Abassi & Papeo, 2020; Papeo et al., 2017; Papeo & Abassi, 2019; Vestner et al., 2019, 2021).

Another unexplored factor is the role of joint attention in dialogues, where listeners anticipate and align with the speaker's focus of attention. Studies in developmental psychology have shown that joint attention is critical for language acquisition and auditory processing (Mundy, 2018; Tomasello, 2005), suggesting that similar mechanisms may also operate in adult conversations to enhance comprehension. Additionally, the contingency of dialogue responses, where listeners predict speaker turns based on context, may activate reward systems (McNaughton et al., 2023; Stivers et al., 2009), further enhancing attention and processing efficiency. Interestingly, the absence of familiar speakers in our study highlights that interactional dynamics alone, independent of familiarity (Johnsrude et al., 2013; Domingo et al., 2020), can drive social context effects. Here, research on speaker familiarity would suggest that combining familiarity with dialogue-based cues could even amplify these benefits (Fleming et al., 2014; Yu et al., 2023). Future studies could investigate how these factors interact to optimize phonetic encoding in real-world scenarios.

An additional point to highlight is that effects of social context were found only with response time, while effects of semantic context were found with both response time and accuracy. A first reason for this difference across tasks may be that our social condition 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 performance on the task and especially measures of accuracy (such as *d'*), that are more sensitive to floor or ceiling effects than RTs (Magnus et al., 2017; Willoughby et al., 2023). Another reason for the discrepancy could be due to the nature of the measured effects themselves. Indeed, it has been proposed that RTs and measures of 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 d-primes, while the exact nature of these different cues, if they exist, would need a more thorough exploration. On a last point, we note that we found the effect of social context to be somewhat more salient when using a design that was mixing all our experimental conditions together within the same presentation blocks as compared to presented in separate blocks. While we don't have a strong interpretation for this phenomenon, we can highlight that effects of experimental design have been anecdotally but regularly reported in the literature (Bock & Griffin, 2000; Carvalho & Goldstone, 2014a, 2014b; Francis & Ciocca, 2003; Hayes & Wedell, 2023; Knock et al., 2000), while lacking a systematic exploration. Thus, we advocate for a more thorough consideration of potential effects of presentation orders prior to designing experimental protocols.

We also note the lack of interaction between semantic and social contexts. Indeed, while both factors independently enhanced phonetic encoding, their effects appeared to operate additively rather than synergistically. This absence of interaction challenges theoretical frameworks suggesting that social and semantic cues are integrated to optimize speech processing (Menenti et al., 2011; Pickering & Garrod, 2021). Here, a potential explanation lies in the modularity of processing streams for semantic and social information. Neural evidence suggests

that semantic processing primarily engages regions in the temporal lobes, such as the angular gyrus and inferior frontal gyrus (Humphries et al., 2007), while social interaction cues involve the STS and medial prefrontal cortex (Hasson et al., 2012; Schilbach et al., 2013). Although these networks are interconnected, their relative independence may explain why semantic and social effects did not interact in our study. It is possible that these processing streams converge at later stages, such as during integration of discourse-level meaning or higher-order inference-making, which were not directly tested in our task. Another possibility is that our task design, which isolated semantic and social factors, may have limited opportunities for these effects to interact. For example, the lack of speaker familiarity or shared conversational goals in our stimuli may have reduced the potential for synergistic effects. Indeed, research on common ground, the shared knowledge between conversational partners, has shown that mutual understanding enhances speech comprehension, particularly in noisy environments (Clark & Brennan, 1991). Future studies could test whether incorporating shared goals or familiar speakers facilitates interaction between semantic and social contexts. Finally, the additive effects observed in our study may reflect the distinct timescales at which semantic and social factors operate. Semantic coherence influences phonetic encoding already at the level of immediate word or sentence processing, while social context effects unfold over longer conversational exchanges. Understanding how these factors converge may require further investigation, particularly using experimental designs that manipulate both immediate and extended contextual cues, or other experimental methods such as neuroimaging.

Another contribution of this study is the exploration of individual differences in social traits, as measured by the AQ instrument, and their relationship with the interplay between social context and phonetic encoding. While the negative relationship between individual AQ scores and

the performance difference between intact and scrambled dialogues was only a trend with a low effect size, these findings suggest that individuals with higher AQ scores may derive less benefit from the social interaction dynamics present in dialogues. This observation is consistent with prior research showing that individuals with higher AQ scores often exhibit reduced sensitivity to social interaction cues, such as joint attention (Bayliss et al., 2005). Moreover, these findings align with previous works showing a negative correlation between AQ scores and performances on measures of sensitivity for other type of social stimuli in the visual modality (Abassi & Papeo, 2022; van Boxtel et al., 2017; Wyer et al., 2012; Yang et al., 2024). This cross-modal pattern raises the possibility of a domain-general mechanism linking social perceptual traits to variation in susceptibility to social cues, whereby reduced sensitivity to social dynamics would manifest similarly across sensory modalities. Such a mechanism would emphasize shared cognitive or neural processes underlying the integration of social cues, regardless of modality, and may reflect fundamental differences in how individuals with higher AQ scores process socially relevant information.

Furthermore, the observed trend doesn't contradict evidence that higher AQ scores are associated with superior attention to fine phonetic detail (Ota et al., 2015; Yu, 2010; Steward & Ota, 2008), but rather suggest a potential trade-off between enhanced sensory discrimination and diminished sensitivity to social cues. Indeed, this dual influence may explain why individuals with higher AQ scores showed less benefit from the social context of dialogues over monologues but could still perform equally well in the monologue condition that emphasized semantic context. Overall, while the current findings provide only weak evidence for this relationship, they highlight the importance of considering individual variability when investigating the interplay between social and semantic contexts in auditory processing. Future research should further examine these

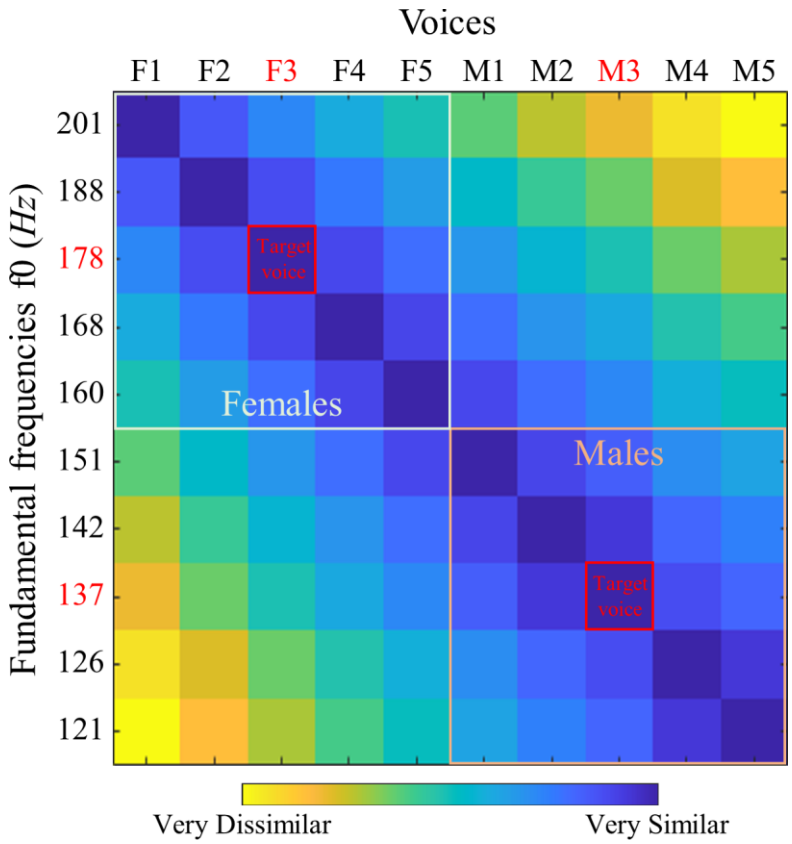
effects using larger samples and refined experimental designs to better elucidate the mechanisms underlying AQ-related differences. Specifically, exploring whether the relationship between AQ or other social traits and social context processing reflects a domain-general mechanism could provide valuable insights into the cognitive architecture of social perception across modalities.

In sum, our findings contribute to several theoretical frameworks in auditory cognition and social perception. First, our results support predictive coding theories, demonstrating that both semantic and social contexts facilitate phonetic encoding, likely through top-down mechanisms. The robustness of these effects across large-scale conversations suggests that predictive mechanisms operate hierarchically, integrating information at multiple levels of linguistic and social processing. Second, our findings highlight the role of social interactional dynamics in engaging cognitive systems specialized for processing socially relevant stimuli, consistent with theories of dialogue-based processing and to neurocognitive models emphasizing the role of specialized neural circuits to social cognition (Pitcher & Ungerleider, 2021). Finally, the trend-level relationship between AQ and sensitivity to social context suggests a balance between enhanced detail-focused processing and reduced integration of social information, aligning with the weak central coherence hypothesis of autism (Happé & Frith, 2006; Hoy et al., 2004; Jolliffe & Baron-Cohen, 1999). However, the variability observed within AQ-related effects here should be understood in the context of a neurotypical population, since we drew our sample from a general population, and none of our participants had a known diagnosis of autism.

These findings also have practical implications for designing interventions and assistive technologies for individuals with hearing losses. For example, auditory training programs could incorporate semantically coherent narratives and simulate conversational dynamics to enhance speech processing in noisy environments. Similarly, hearing aids and other assistive devices could

incorporate algorithms that prioritize interactional dynamics, such as turn-taking, to improve real-world communication outcomes. Future research should extend the examination of the neural mechanisms underlying these effects using neuroimaging techniques to explore how different contexts modulate brain activity (see Landsiedel & Koldewyn, 2023; Olson et al., 2023). On another aspect, expanding the scope of this research to include older adults, individuals with hearing impairments, or non-native speakers could provide a broader understanding of how contextual factors impact diverse populations. Finally, integrating cross-modal research on auditory and visual sensitivity to social interactions (see Abassi et al., 2024 for a review on social vision; see also Abassi & Papeo, 2024 for cross category research) could offer a more comprehensive perspective on disentangling the domain-general and domain-specific mechanisms underlying social cognition.

IX. 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 babble noise.

a. d'

Info	With Baseline	Without Baseline	Δ	Comment
LogLikelihood	-955	-985	30.62	
AIC	1950	2009	-59.23	Less is better
BIC	2044	2099	-54.49	Less is better

b. RTs

Info	With Baseline	Without Baseline	Δ	Comment
LogLikelihood	22.2	-31.8	53.98	
AIC	-4.4	101.6	-105.96	Less is better
BIC	90.36	191.6	-101.23	Less is better

Supplementary Table 1. Comparison of Linear Mixed Effect models fits with and without including the baseline as a covariate.

a. d'

	F	df	df (res)	p
Design	2.33	2.00	208.00	0.100
Social	1.61	1.00	398.00	0.205
Semantic	7.35	1.00	210.00	**0.007**
Design x Social	0.05	2.00	398.00	0.956
Design x Semantic	0.40	2.00	210.00	0.670
Social x Semantic	1.74	1.00	416.00	0.188
Design x Social x Semantic	0.38	2.00	416.00	0.681

b. RTs

	F	df	df (res)	p
Design	0.82	2	208	0.440
Social	4.40	1	208	**0.037**
Semantic	9.55	1	208	**0.002**
Design x Social	2.84	2	208	0.061
Design x Semantic	1.59	2	208	0.206
Social x Semantic	0.02	1	208	0.889
Design x Social x Semantic	0.75	2	208	0.472

Supplementary Table 2. Results of the Linear Mixed Effect models with d' (a.) and RTs (b.) for Experiment 2 without baseline as a covariate.

X. ACKNOWLEDGMENTS. This research was supported by the Fondation Pour l'Audition through a postdoctoral fellowship to EA (FPA RD-2022-1) and by the Grand Prix Scientifique from the Fondation Pour l'Audition to RJZ (2021). RJZ holds a Canada Research Chair and is a fellow of the Canadian Institute for Advanced Research.

XI. AUTHOR DECLARATIONS. All authors declare that they have no financial or non-financial conflict of interests to disclose. They also declare that all procedures used in the current study were performed in compliance with the code of ethics of the World Medical Association's Declaration of Helsinki and have been approved by the Faculty of Medicine Institutional Review Board of McGill University on the 3rd of October 2022 (reference number: A08-B75-22B / 22-07-116).

XII. DATA AVAILIBILITY. The entire set of stimuli used in the current study can be found online at www.zlab.mcgill.ca/stimuli/ ('Stimuli_from_Abassi_2025') and can be used under the

term of a CC-BY-NC-SA 4.0 license. The collected data are freely available upon email request to E.A or R.Z.

XIII. REFERENCES

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

Abassi, E., & Papeo, L. (2022). Behavioral and neural markers of visual configural processing in social scene perception. *NeuroImage*, 260. https://doi.org/10.1016/j.neuroimage.2022.119506

Abassi, E., & Papeo, L. (2024). Category-selective representation of relationships in visual cortex. *Journal of Neuroscience*, 44(5). https://doi.org/10.1523/JNEUROSCI.0250-23.2023

Abassi, E., Bognár, A., de Gelder, B., Giese, M., Isik, L., Lappe, A., Mukovskiy, A., Solanas, M.P., Taubert, J. & Vogels, R., (2024). Neural Encoding of Bodies for Primate Social Perception. *Journal of Neuroscience*, *44*(40). https://doi.org/10.1523/JNEUROSCI.1221-24.2024

Adolphs, R. (2001). The neurobiology of social cognition. *Current opinion in neurobiology*, 11(2), 231-239. https://doi.org/10.1016/s0959-4388(00)00202-6

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

Babel, M., & Russell, J. (2015). Expectations and speech intelligibility. *The Journal of the Acoustical Society of America*, *137*(5), 2823–2833. https://doi.org/10.1121/1.4919317

Baker, R., & Hazan, V. (2010). LUCID: a corpus of spontaneous and read clear speech in British English. *Proc. DiSS-LPSS Joint Workshop (DiSS 2010)*, 3-6

Baker, R., & Hazan, V. (2011). DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, *43*(3), 761–770. https://doi.org/10.3758/S13428-011-0075-Y

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

Bayliss, A. P., Pellegrino, G. Di, & Tipper, S. P. (2005). Sex differences in eye gaze and symbolic cueing of attention. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *58*(4), 631–650. https://doi.org/10.1080/02724980443000124

Beechey, T. (2022). On Perspective Taking in Conversation and in Research: A Comment on Bambara et al. (2021). *Journal of Speech, Language, and Hearing Research*, *65*(4), 1597–1599. https://doi.org/10.1044/2021_JSLHR-21-00606

Belin, P., Zatorre, R. J., Lafallie, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, 403, 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.

Bock, K., & Griffin, Z. M. (2000). The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General*, *129*(2), 177–192. https://doi.org/10.1037/0096-3445.129.2.177

Brandman, T., & Peelen, X. M. V. (2017). Interaction between scene and object processing revealed by human fMRI and MEG decoding. *Journal of Neuroscience*, 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

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, Perception, and Psychophysics*, *77*(5), 1465–1487. https://doi.org/10.3758/S13414-015-0882-9

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

Carvalho, P. F., & Goldstone, R. L. (2014a). Effects of Interleaved and Blocked Study on Delayed Test of Category Learning Generalization. *Frontiers in Psychology*, *5*(AUG), 87396. https://doi.org/10.3389/FPSYG.2014.00936

Carvalho, P. F., & Goldstone, R. L. (2014b). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin and Review*, *22*(1), 281–288. https://doi.org/10.3758/S13423-014-0676-4

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

Clements, G. N. (1985). The geometry of phonological features. Phonology, 2, 225-252. https://doi.org/10.1017/S0952675700000440

Coffey, E. B. J., Arseneau-Bruneau, I., Zhang, X., & Zatorre, R. J. (2019). The Music-In-Noise Task (MINT): A Tool for Dissecting Complex Auditory Perception. *Frontiers in Neuroscience*, *13*, 199. https://doi.org/10.3389/FNINS.2019.00199

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

Cox, R. M., Alexander, G. C., & Gilmore, C. (1987). Development of the Connected Speech Test (CST). *Ear and Hearing*, *8*(5 Suppl), 119S-126S. https://doi.org/10.1097/00003446-198710001-00010

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

Daltrozzo, J., Signoret, C., Tillmann, B., & Perrin, F. (2011). Subliminal Semantic Priming in Speech. *PLOS ONE*, *6*(5), e20273. https://doi.org/10.1371/JOURNAL.PONE.0020273

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

DeWitt, I., & Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(8), 2709. https://doi.org/10.1073/PNAS.1113427109

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

Domingo, Y., Holmes, E., & Johnsrude, I. S. (2020). The benefit to speech intelligibility of hearing a familiar voice. *Journal of Experimental Psychology: Applied*, *26*(2), 236. https://doi.org/10.1037/xap0000247

Drager, K., & Kirtley, M. J. (2016). Awareness, salience, and stereotypes in exemplar-based models of speech production and perception. *Awareness and Control in Sociolinguistic Research*, 1–24.

Farbood, M. M., Heeger, D. J., Marcus, G., Hasson, U., & Lerner, Y. (2015). The neural processing of hierarchical structure in music and speech at different timescales. *Frontiers in Neuroscience*, *9*(APR), 132041. https://doi.org/10.3389/FNINS.2015.00157

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

Fleming, D., Giordano, B. L., Caldara, R., & Belin, P. (2014). A language-familiarity effect for speaker discrimination without comprehension. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(38), 13795–13798. https://doi.org/10.1073/PNAS.1401383111

Foss, D. J. (1982). A discourse on semantic priming. *Cognitive Psychology*, *14*(4), 590–607. https://doi.org/10.1016/0010-0285(82)90020-2

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

Francis, A. L., & Ciocca, V. (2003). Stimulus presentation order and the perception of lexical tones in Cantonese. *The Journal of the Acoustical Society of America*, *114*(3), 1611–1621. https://doi.org/10.1121/1.1603231

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

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

Happé, F., & Frith, U. (2006). The weak coherence account: Detail-focused cognitive style in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *36*(1), 5–25. https://doi.org/10.1007/S10803-005-0039-0

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: A mechanism for creating and sharing a social world. Trends in Cognitive Sciences, 16(2), 114–121. https://doi.org/10.1016/j.tics.2011.12.007

Hautus, M. J., Macmillan, N. A., & Creelman, C. D. (2021). *Detection theory: A user's guide*. Routledge.

Hayes, W. M., & Wedell, D. H. (2023). Effects of blocked versus interleaved training on relative value learning. *Psychonomic Bulletin and Review*, *30*(5), 1895–1907. https://doi.org/10.3758/S13423-023-02290-6

Hoy, J. A., Hatton, C., & Hare, D. (2004). Weak Central Coherence: A Cross-Domain Phenomenon Specific to Autism? *Autism*, 8(3) 267–281. https://doi.org/10.1177/1362361304045218

Humphries, C., Binder, J. R., Medler, D. A., & Liebenthal, E. (2007). Time course of semantic processes during sentence comprehension: an fMRI study. Neuroimage, 36(3), 924-932. https://doi.org/10.1016/j.neuroimage.2007.03.059

Iamshchinina, P., Karapetian, A., Kaiser, D., & Cichy, R. M. (2022). Resolving the time course of visual and auditory object categorization. *Journal of Neurophysiology*, *127*(6), 1622-1628. https://doi.org/10.1152/jn.00515.2021

Jakobson, R., Fant, C. G., & Halle, M. (1951). Preliminaries to speech analysis: The distinctive features and their correlates. https://hdl.handle.net/11858/00-001M-0000-002B-AD6A-3

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological science*, *24*(10), 1995-2004. https://doi.org/10.1177/095679761348246

Jolliffe, T., & Baron-Cohen, S. (1999). A test of central coherence theory: linguistic processing in high-functioning adults with autism or Asperger syndrome: is local coherence impaired? *Cognition*, *71*(2), 149–185. https://doi.org/10.1016/S0010-0277(99)00022-0

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

Knoblauch, K. (2023). *psyphy: Functions for Analyzing Psychophysical Data in R*. https://CRAN.R-project.org/package=psyphy

Knock, T. R., Ballard, K. J., Robin, D. A., & Schmidt, R. A. (2000). Influence of order of stimulus presentation on speech motor learning: A principled approach to treatment for apraxia of speech. *Aphasiology*, *14*(5–6), 653–668. https://doi.org/10.1080/026870300401379

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

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*(Volume 62, 2011), 621–647. https://doi.org/10.1146/ANNUREV.PSYCH.093008.131123

Lahiri, A., & Reetz, H. (2010). Distinctive features: Phonological underspecification in representation and processing. *Journal of Phonetics*, *38*(1), 44-59. https://doi.org/10.1016/j.wocn.2010.01.002

Landsiedel, J., & Koldewyn, K. (2023). Auditory dyadic interactions through the “eye” of the social brain: How visual is the posterior STS interaction region? *Imaging Neuroscience*, *1*, 1–20. https://doi.org/10.1162/imag_a_00003

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

Lenhard, W., & Lenhard, A. (2014). *Hypothesis Tests for Comparing Correlations*. https://doi.org/10.13140/RG.2.1.2954.1367

Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, *46*(6), 1093–1096. https://doi.org/10.1016/J.JESP.2010.05.025

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. *Ninth International Conference on Spoken Language Processing (ICSLP)*, Pittsburgh, United States. 17 - 21 Sep 2006. http://eprints.soton.ac.uk/id/eprint/188337

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. *Assessment* 26(7), 1296-1306. https://doi.org/10.1177/1073191117740953

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

McGowan, K. B. (2015). Social Expectation Improves Speech Perception in Noise. *Language and Speech*, *58*(4), 502–521. https://doi.org/10.1177/0023830914565191

McNaughton, K.A., Kirby, L.A., Warnell, K.R., Alkire, D., Merchant, J.S., Moraczewski, D., Yarger, H.A., Thurm, A. & Redcay, E. (2023). Social-interactive reward elicits similar neural response in autism and typical development and predicts future social experiences. *Developmental Cognitive Neuroscience*, *59*, 101197. https://doi.org/10.1016/j.dcn.2023.101197

Menenti, L., Garrod, S. C., & Pickering, M. J. (2012). Toward a neural basis of interactive alignment in conversation. *Frontiers in Human Neuroscience*, *6*, 24405. https://doi.org/10.3389/FNHUM.2012.00185

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

Mundy, P. (2018). A review of joint attention and social-cognitive brain systems in typical development and autism spectrum disorder. *European Journal of Neuroscience*, *47*(6), 497–514. https://doi.org/10.1111/EJN.13720

Niedzielski, N. (1999). The Effect of Social Information on the Perception of Sociolinguistic Variables. *Journal of language and social psychology, 18(1), 62-85*. https://doi.org/10.1177/0261927X99018001005

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

Ota, M., Stewart, M. E., Petrou, A. M., & Dickie, C. (2015). Lexical effects on children's speech processing: individual differences reflected in the autism-spectrum quotient (AQ). Journal of Speech, Language, and Hearing Research, 58(2), 422-433. https://doi.org/10.1044/2015_JSLHR-L-14-0061

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

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

Pickering, M. J., & Garrod, S. (2021). Understanding Dialogue: Language Use and Social Interaction. *Cambridge University Press*. ISBN: 9781108461931 https://philpapers.org/rec/PICUDL-2

Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a third visual pathway specialized for social perception. Trends in cognitive sciences, 25(2), 100-110. https://doi.org/10.1016/j.tics.2020.11.006

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

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/

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

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education*, *33*(4), 511–531. https://doi.org/10.1007/BF00973770

Sapir, E. (1925). Sound patterns in language. *Language*, *1*(2), 37-51. https://doi.org/10.2307/409004

Schilbach, L., Eickhoff, S. B., Rotarska-Jagiela, A., Fink, G. R., & Vogeley, K. (2008). Minds at rest? Social cognition as the default mode of cognizing and its putative relationship to the "default system" of the brain. *Consciousness and Cognition*, *17*(2), 457–467. https://doi.org/10.1016/J.CONCOG.2008.03.013

Schilbach, L., Wohlschlaeger, A. M., Kraemer, N. C., Newen, A., Shah, N. J., Fink, G. R., & Vogeley, K. (2006). Being with virtual others: Neural correlates of social interaction. *Neuropsychologia, 44(5), 718-730*. https://doi.org/10.1016/j.neuropsychologia.2005.07.017

Schilbach, L., Eickhoff, S. B., Rotarska-Jagiela, A., Fink, G. R., & Vogeley, K. (2013). Minds at rest? Social cognition as the default mode of cognizing and its putative relationship to the "default system" of the brain. Consciousness and Cognition, 21(1), 133–143. https://doi.org/10.1016/j.concog.2011.12.013

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

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

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K. E., & Levinson, S. C. (2009). Universals and cultural

variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(26), 10587–10592. https://doi.org/10.1073/PNAS.0903616106

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

Tomasello, M. (2005). *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.

Tóth, B., Honbolygó, F., Szalárdy, O., Orosz, G., Farkas, D., & Winkler, I. (2020). The effects of speech processing units on auditory stream segregation and selective attention in a multi-talker (cocktail party) situation. *Cortex*, *130*, 387–400. https://doi.org/10.1016/J.CORTEX.2020.06.007

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

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

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

Yang, G., Wang, Y., & Jiang, Y. (2024). Social perception of animacy: Preferential attentional orienting to animals links with autistic traits. *Cognition*, *251*, 105900. https://doi.org/10.1016/J.COGNITION.2024.105900

Yi, H. G., Leonard, M. K., & Chang, E. F. (2019). The encoding of speech sounds in the superior temporal gyrus. *Neuron*, *102*(6), 1096-1110. https://doi.org/10.1016/j.neuron.2019.04.023

Yu, A. C. L. (2010). Perceptual Compensation Is Correlated with Individuals' "Autistic" Traits: Implications for Models of Sound Change. *PLOS ONE*, *5*(8), e11950. https://doi.org/10.1371/journal.pone.0011950

Yu, K., Zhou, Y., Zhang, L., Li, L., Li, P., & Wang, R. (2023). How Different Types of Linguistic Information Impact Voice Perception: Evidence From the Language-Familiarity Effect. *Language and Speech*, *66*(4), 1007–1029. https://doi.org/10.1177/00238309221143062

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