Science Advances
AAAS

Supplementary Materials for

The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation

Andrew Reece *et al.*

Corresponding author: Andrew Reece, andrew.reece@betterup.com; Gus Cooney, guscooney@gmail.com

Sci. Adv. **9**, eadf3197 (2023)
DOI: 10.1126/sciadv.adf3197

This PDF file includes:

SUPPLEMENTARY MATERIALS

RECRUITMENT AND DEMOGRAPHIC INFORMATION

Table S.1. *Dates and Sample Size of Each Data Collection Round. N Unique Speakers Reflects the Number of Unique Speakers per Round. Each Conversation Had Two Speakers.*

Recruitment Round	Recruitment Date	*N* Conversations	Percent of Conversations	*N* Unique Speakers
Round 1	01/07 - 01/14	46	2.78	92
Round 2	05/01 - 05/29	183	11.05	366
Round 3	06/26 - 07/14	196	11.84	391
Round 4	07/29 - 08/31	403	24.34	475
Round 5	10/08 - 11/07	423	25.54	480
Round 6	11/09 - 11/25	405	24.46	493

Table S.2. *Demographic Information for Participants in the Corpus.*

Demographics		Sample *N*	Sample Percent
Age	18-25	425	29.19
	25-35	499	34.27
	35-45	286	19.64
	45-55	129	8.86
	55+	83	5.7
	Not Reported	34	2.34
Gender	Female	782	53.71
	Male	610	41.9
	Other or Prefer not to Answer	30	2.06
	Not Reported	34	2.34
Race/Ethnicity	White	920	63.19
	Asian	200	13.74
	Black or African American	117	8.04
	Hispanic or Latino	108	7.42

Demographics		Sample *N*	Sample Percent
	Mixed	53	3.64
	Other	13	0.89
	American Indian or Alaska Native	7	0.48
	Native Hawaiian or Pacific Islander	2	0.14
	Prefer not to Say	2	0.14
	Not Reported	34	2.34
Education	Bachelor's Degree	567	38.94
	Some College	354	24.31
	Master's Degree	247	16.96
	Associate Degree	97	6.66
	Completed High School	81	5.56
	Professional Degree	36	2.47
	Doctoral Degree	32	2.2
	Some High School	8	0.55
	Not Reported	34	2.34

Note. N = 1456.

CORPUS CONSTRUCTION

Data Processing & Feature Extraction

Here, we describe in detail how the video and audio files were processed into unified, structured, and user-friendly formats.

Conversational Alignment

Each participant's video stream was saved as an independent video in .mkv format. If a participant's connection dropped and then rejoined the conversation session, a new video file was created in addition to the existing one. As such, processing of recorded conversations started with the creation of a coherent, single-file representation of the conversation from each partner's respective video files. Programmatic alignment consisted of four primary steps using the video processing software, FFMPEG. First, input media were reencoded to correct possibly corrupted timestamps. Second, the TokBox metadata, which provided a timeline of when participants joined, left, and possibly rejoined the conversation, was verified and corrected by measuring the

duration of the media. Third, after metadata correction, an individual participant's videos were combined into a single video, adding padding and blank filler segments where appropriate. For example, if a participant dropped and rejoined 10 seconds later, two videos would have been created in the archive. Directly joining these videos was not helpful, because of that 10 second gap; the resulting merged video would be misaligned. However, the gap in time was also reflected in the metadata, and so a 10 second "blank" video was inserted to maintain alignment with respect to the overall conversational timeline. Finally, participants' aligned videos were combined into a unified representation. Separate audio channels for each participant were created for downstream automated transcription.

It is worth noting that for some video segments the TokBox software does not accurately record the start and stop time of the video stream correctly relative to the overall timeline of the conversation. Video durations (and therefore offsets) can be verified using the FFMPEG tool FFPROBE to measure the audio and video stream durations and compare them to those reported in the metadata, adding or subtracting appropriate offsets to the stopTimeOffset where necessary. Correcting the startTimeOffset is more difficult and requires heuristics since there is no trusted reference point in time. We chose the heuristic of minimizing audio signal overlaps during playback as a proxy for proper alignment. Such a heuristic is imperfect which impacts a small number of conversations in the dataset leading to slight misalignment in the unified audio signals; the independent signals are also included. Finally, note that videos where the alignment was problematic in human review were excluded.

Overall, joining the conversation videos posed a *non-trivial* challenge, and required a number of subjective, albeit carefully reasoned, decisions. The corpus therefore includes both the

raw video files along with their merged versions, so that other researchers may apply their own alternate methods for alignment and synchronization.

Feature Extraction

We describe the processes used to generate and extract analyzable features according to the source of the information: transcripts, audio, video, and survey responses. The outputs of these processes are analysis-ready in the sense that they are structured into common file formats and indexed by a shared timeline or by conversational turns where appropriate.

Textual. Processing of textual information involved transcription, turn identification, and extracting speaking statistics.

Transcription. To produce a transcript of the conversation, we processed the aligned conversation files using the Amazon Web Services (AWS) *Transcribe* automated transcription service. The raw transcript and tokens returned by the Transcribe API are included in this data release.

Note that while the transcripts are very usable, the quality of automated transcription is far from perfect. Throughout the development of this dataset, we tested numerous automated transcription services, and each of them left much to be desired. An important direction for future work on this dataset is the development of "gold standard" transcripts, either via improved automated transcription or human labeling.

Turn identification. The sequential representation of text in alternating turns is essential to many conversation analyses. The definition of a turn, however, can vary, depending on how pauses, overlaps, backchannels, and other complications are preprocessed. The simplest way to construct a conversational turn is to assign each word token to a participant's turn until a token from another participant occurs, at which point that participant's turn begins, and so on. This is

the default method used to construct turns from the raw transcript. Although limited in many respects, this approach provides a useful reference point for improved algorithms that we develop (see Section 1 of the Results in the Manuscript).

Turn-based and time-based feature aggregation. Once equipped with conversational turns, we considered two possible approaches to aggregating the remaining acoustic, visual, and textual features: time-based (finer-grained) and turn-based (coarser). In the release, we include time-based aggregations at a one-second resolution. Researchers can use the turn timestamps noted in the transcript files to aggregate turn-based features as desired. For a comprehensive list of corpus features and how they were computed, see the Data Dictionary.

Note that while the entire TokBox session is included for each conversation, the conversation is said to have begun the moment both participants have joined the session. In all of the turn-based indexing included in this release, this moment is specified as turn_id=0, with all prior data for the session being indexed as turn_id=-1. So, if you are watching a conversation video and observe a particular moment of interest, you could locate the turn number by searching the utterance field of the turn-based aggregation of the conversation. Using the conversation ID and turn number, you could then index into any of the other features discussed in this section; for example, the probability that each participant is displaying a happy facial expression during that turn.

Acoustic. Processing of acoustic information involved spectral characterizations, phonation, and prosody features.

Spectral characterizations. A number of calculations were performed on the audio of the conversations. For example, the fundamental frequency (F0) of people's speech was computed over 0.01 s intervals using the Parselmouth package (102). Further, the Python library Librosa

(103) was used to compute the first 13 Mel-frequency cepstrum coefficients, as well as various additional spectral features such as the spectral centroid, contrasts, zero-crossing rate, and others (104). These features were aggregated using the mean value over 1 s intervals.

Phonation and prosody. Baseline vocal pitch (F0) was used along with signal energy to compute other prosodic features; for example, jitter and shimmer, which measure the variance of pitch and volume respectively (105). We also computed a measure of vocal "intensity" (sometimes referred to as "activation"), which is a measure of emotion and momentary affect (See Section 4 of the Results in the Manuscript). To do so, we trained a model on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (82), and then applied this model to our corpus. These features were aggregated using the mean value over 1 s intervals.

Visual. Processing of visual information involved smile, nod, and emotion detection. Visual features were computed at 1 s intervals (frames). In some frames, a face was not detected, and for these frames the visual features were recorded as null values.

Smile and nod. We used OpenCV, a computer vision software library, to extract a set of facial landmark locations within each visual frame. We then applied a set of heuristics to estimate whether or not a participant was smiling or nodding for any given frame. We attempted to measure gaze as well but were not satisfied with the consistency of results.

Emotion detection. An emotion recognition model was trained using the AffectNet dataset (84). The model, a convolutional neural network, assigned a probability distribution across eight emotional classes (happy, sad, angry, etc.) to each frame of facial expression, per speaker.

CORPUS FRAMEWORK

Our “levels” framework simply helps to organize a vast, multi-featured dataset into convenient categories for reporting analyses that clearly belong to different families of content. Here we discuss in more detail how this notion of a conversational hierarchy, in addition to its practical utility, may also prove fruitful in generating new theoretical insights.

Consider a lower-level conversational feature as a purely descriptive statement of record. The pitch of a speaker’s voice, the presence or absence of eye contact, and time spent in silence are all examples of low-level features. Typically, these actions can be captured at a sub-second timescale. While it is true that the automated extraction of such features from a recording may require considerable feats of algorithmic inference (e.g., tracking on-screen gaze), these features nevertheless at least seek to capture an objective record of a conversation.

Moving upward through the hierarchy, different levels are characterized by the degree of indirect inference required and the breadth of contextual information used to make these inferences. The distinction we make between a mid-level versus a high-level inference therefore becomes a matter of scope. Humans, it seems, often employ a wide range of inputs to make their judgments (e.g., Whether Jill likes Jack is a highly indirect inference based on information across space, time, and textual/acoustic/visual modalities). Moreover, humans frequently, perhaps even rather helplessly, employ the full scope of their lived experiences to make sense of the present moment (e.g., “This person’s voice reminds me of my dear Aunt Sally, whom I remember fondly”). Because subjective impressions and judgments about a conversation incorporate the broadest range of information and context, we distinguish them as *high-level* inferences.

In contrast, mid-level inferences are characterized by their use of a narrower scope of context and antecedent reference. Informally, they may dig deep, but not wide, to know what

they know. For example, language embeddings, which create numerical representations of the semantic meaning of spoken words (See Section 4 of the Results in the Manuscript), are made possible because an underlying statistical model was trained on a deep and extensive corpus of written language. While this kind of inference is based on a wealth of previously encountered information—similar to human judgment, in that regard—the context behind language embeddings is arguably deeper (i.e., billions of words of text training data) than it is wide. A hitch in the voice, a sad glance away; all these signals, essential for shared understanding between humans, will go unnoticed by a machine that knows only language. We thus refer to these inferences, which typically vary on the timescale of a conversational turn, as mid-level inferences. This layer of the conversational hierarchy thus operates somewhere between the objective immediacy of low-level events and the subjective expanse of high-level human judgment.

Notably, humans, too, can make mid-level inferences in conversation, such as when some aspect of the conversation carries a particular salience (e.g., a captivating facial expression), at which point people often stop attending to the wider array of signals that normally influence their ongoing impression formation. Similarly, but in the converse, the more algorithms are able to account for context that once seemed solely in the domain of human capacity, the more they seem eerily human, threatening to cross over the safety of the uncanny valley. Despite the crossing of levels by human and machine, discretizing this continuum of inference into a middle and higher level seems of theoretical and practical utility.

One unresolved question in conversation research relates to how information flows *across* levels. It is clear that low-level, factual accounting of a conversation must necessarily

underlie higher-order sense-making about a conversation. From there, however, the trajectory of inference remains an open area for future study.

One possible representation of the hierarchy is as a cascade of dependencies, with high-level judgments relying on mid-level inferences, which in turn draw on low-level behaviors to convey meaning. We invoked this model in the main text, with an example of a smile (low-level), which may be a key component of what is perceived as a happy facial expression (mid-level), which in turn may serve as input into one's assessment of their conversation as enjoyable (high-level). But it seems equally plausible that low-level features may influence high-level impressions directly, without moving "through" a mid-level inference.

To complicate matters further, once formed, high-level impressions may percolate back down through the hierarchy. For example, after registering a perceived insult, a speaker's low-level behavior may change—elevating vocal pitch, increasing facial tension, or clipping speech. In turn, subsequent mid-level perceptions may be distorted, leading participants to draw differing conclusions about the same objective events. Scholars of conversation currently have little empirical basis for choosing among these intuitively plausible models for information flow, and their resistance to a simple accounting reflects the complexity of human conversation. Ultimately, information dynamics within these levels of conversation remains a subject in need of considerable future research.

As this discussion makes clear, the results we present only scratch the surface of the corpus. We regard our findings and this simply framework as an initial overture that will require additional efforts from many researchers across the social and computational science. Scholars from a variety of disciplines appear increasingly interested in the dynamics of conversation, and

there are countless aspects of conversation that we did not begin to cover. The raw material is there, however—and in many cases, processed and ready to be analyzed.

RESULTS

Section on Turn-taking and Turn Segmentation

Turn Exchange

Conversations were transcribed with AWS Transcribe. Each transcription's most basic form was a list of individual tokens, accompanied by start and end timestamps, speaker IDs, and confidence estimates. The minimum temporal resolution was 10 ms (0.01 s). Within each conversation and speaker, tokens were joined with adjacent tokens if 20ms or less of pause separated them. This output is considered the Heldner & Edlund transcript, where each row is considered a speaking turn.

Following this, the Heldner & Edlund (35) communication state classification algorithm was applied to each transcript. This algorithm created, for each conversation, a time-series at 10ms increments of who, if anyone, was speaking at that moment in time. Using this time-series, a new dataset was created where each state transition was classified as either a Gap (between-speaker silence), Pause (within-speaker silence), Overlap (between-speaker overlap), or WSO (within-speaker overlap, an interruption). The units for each classification are durations in milliseconds. Overlaps are the only intervals which have values below 0 ms.

To address outliers, we removed between- and within-speaker intervals more than three standard deviations from the mean. Pauses and WSOs cannot be less than 0 s. Accordingly, we removed outliers >3 SD above the mean in the case of these two measures. Upon visual inspection, we observed that these outliers were nearly always attributable to technical issues, such as moments of poor internet connectivity, rather than genuine conversational anomalies.

Turn Duration

```
TERMINAL_PUNC_CUES = [
    ".",
    "?",
    "!",
]
```

Backchannels

```
backchannel_CUES = [
    "a",
    "ah",
    "alright",
    "awesome",
    "cool",
    "dope",
    "e",
    "exactly",
    "god",
    "gotcha",
    "huh",
    "hmm",
    "mhm",
    "mm",
    "mmm",
    "nice",
    "oh",
    "okay",
    "really",
    "right",
    "sick",
    "sucks",
    "sure",
    "uh",
    "um",
    "wow",
    "yeah",
    "yep",
    "yes",
    "yup",
]

NOT_backchannel_CUES = [
    "and",
    "but",
```

```
    "i",
    "i'm",
    "it",
    "it's",
    "like",
    "so",
    "that",
    "that's",
    "we",
    "we're",
    "well",
    "you",
    "you're",
]
```

Section on Conversation and Wellbeing

We were encouraged during the review process to further explore the effect of repeatedly answering questions about one's wellbeing. While we can never entirely rule out that repeated rating accounts for some part of the effect, we can further examine our data for evidence that intervention itself (i.e., the conversation) is, in fact, the primary driver of the observed improvement in people's subjective well-being.

To this end, we believe that some kind of demand effect offers the most plausible explanation for why repeated measurement itself might have caused an increase in self-reported well-being. That is, participants reported higher well-being simply because they were asked the same question twice and assumed that we expected them to provide a higher score. We therefore examine some observable implications of these competing explanations for the increase.

Consider that if the pre-and-post-conversation changes that we observed were simply due to the demand of asking the question twice, one prediction of this account would be that the magnitude of the change would *not* depend on the quality of the conversation that people had with their partner. To examine this, we divided our sample into quartiles. We then analyzed data from participants who reported that their conversation partners were in the bottom quartile of

conversational quality (i.e., “Bad Conversation Partners”) and the top quartile (i.e., “Good Conversation Partners”). A model with conversation quality added as moderator revealed that conversation partner quality significantly moderated the size of people’s pre-post affect change (b = 0.56, 95% CI = [0.40, 0.72], $t(1624) = 6.74$, $p < .001$).

We also performed this analysis using a more ‘objective’ measure of conversation quality: the length/duration of people’s conversations. To examine this, we again divided our sample of conversations into quartiles based on duration. We then analyzed data from participants whose conversations were in the bottom quartile of duration (i.e., “Short Conversations”) and in the top quartile (i.e., “Long Conversations”). A model with conversation duration added as moderator revealed that duration significantly moderated the size of people’s pre-post affect change (b = 0.84, 95% CI = [0.69, 0.98], $t(1919) = 11.28$, $p < .001$).

In short, people who reported talking to better conversation partners and those who had longer conversations had *larger* pre-post affect changes compared to those who had worse conversation partners and shorter conversations. As a result, the weight of the evidence appears inconsistent with the idea that the pre-post affect change we observed is the result of asking the question twice.

While these additional analyses do not completely rule out the presence of demand, they do show that a demand effect cannot account for the entirety of our effect. People’s post-conversation reported well-being is not just a function of being asked a question twice, but it is also a function of the quality of the conversation (as measured by people’s self-reported partner quality and the objective length of the conversation). The conversation itself is acting as significant input into people’s well-being.

Section on Good Conversationalists

Section S.1 begins by explaining how we computed turn-level audio, visual, and textual features. In S.2, we then describe our statistical procedure for assessing differences on these turn-level features among participants who varied in their partner-rated conversationalist scores; we also describe our procedure for multiple-testing adjustment of *p* values. Section S.3 presents complete results for our analysis of partner-rated conversationalist scores, including (a) additional features not discussed in the main text; (b) patterns of results for the "middling" conversationalist groups, defined as those in the 25-50th and 50-75th percentiles of their partner-rated score; and (c) gender-specific and turn-duration-adjusted results.

S.1. Obtaining Turn-level Features

In this section, we describe how continuous audio recordings and image frames were aggregated, based on start/stop times in a transcript segmented by speaking turns, into turn-level summary features describing speaker and listener behavior. We also describe how transcript-based summary features were extracted from a segmented transcript. We used the Backbiter turn segmentation model, although the same procedure could be employed with any speaker-attributed transcript with turn start/stop timestamps (for a discussion of turn segmentation, see Section 1 of the Results in the Manuscript).

S.1.1. Transcript Features

Six transcript-based features were computed for each turn. Of these, four were straightforward summary statistics extracted from the segmented transcript.

- *Pause*: The difference between the end time of the prior turn and the start time of the current turn. This value is negative for turns that overlap with the previous speaker and positive for those that are preceded by a period of silence.
- *Duration*: The difference between the start and end times of the current turn.
- *Speech rate:* The number of words uttered during a turn, divided by the turn duration.
- *Backchannel rate*: The number of backchanneling events by a listener during a turn (See Backchannel section in Section 1 of the Results in the Manuscript), divided by the turn duration.

The remaining two textual features use pre-trained sentence embedding models to convert turn-level transcripts into a vector representation. Our main results are based on a sentence-level embeddings implementation (73) of MPNet (72). We used cosine similarity and Euclidean distance as two measures of semantic distance.

A common alternative measure of semantic distance, the dot product between two vectors, is identical to cosine similarity in our application because MPNet results are standardized to unit length. To evaluate the robustness of our results to different embedding models, Section S.3 reports comparative results using RoBERTa embeddings (74).

S.1.1. Audio Features

Six audio features are reported for each turn, aggregating a variety of lower-level measures computed at various timescales: short-term (corresponding to 40-millisecond intervals), medium-term (1 s intervals) and long-term (speaker turns of varying length). This aggregation proceeds in two steps.

First, short-term values for numerous low-level features were computed by summarizing the audio signal in rolling 40-millisecond windows. These low-level features included whether the window contained voiced speech, the fundamental vocal frequency of that speech (F0, measured in Hz), volume (log energy, proportional to decibels), and 14 Mel-frequency cepstral coefficients (MFCCs) that describe the shape of the power spectrum. Low-level feature extraction was conducted using the Python libraries librosa (103), Parselmouth (102; 106), pysptk (107) and DisVoice (108).

These short-term auditory measures were aggregated at 1 s resolution. Average pitch and loudness were computed by taking the average non-missing values of all frames within a 1 s interval. This aggregation step helped address peculiarities in certain audio features, such as the fact that the fundamental frequency is undefined in windows of unvoiced speech (e.g., during the unvoiced sibilant /s/). In addition to these transparent averages of objective speech attributes, we also computed two model-based proxies of emotional expression—concepts which can be difficult to directly measure due to the subjective nature of their perception.

Because human annotation of speech is highly labor intensive, we utilized labeled data from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS; 82) to train a computational model that was subsequently applied to our corpus. The RAVDESS dataset contains recordings of 24 trained actors reading statements of varying emotional categories (e.g., calm, happy, sad, angry, fearful, surprise, disgust, or neutrality), expressed with either normal or high emotional intensity. To estimate speech intensity, we computed a series of summary statistics—mean, maximum, and standard deviation for fundamental frequency, log energy, and voiced and unvoiced duration—for the short-term feature time-series within each 1 s interval in our corpus. Medium-term summary statistics were then input into a logistic regression trained on

intensity labels and similarly featurized 1 s intervals from the RAVDESS corpus (described below). The resulting model predictions were used as a proxy for speech intensity in our corpus.

Finally, five medium-term measures were aggregated to long-term turn measures as follows.

- *Pitch*: Average of 1 s fundamental frequency values from speaker audio channel within each turn (includes all 1 s intervals from turn start to end)
- *Pitch variation*: Standard deviation of 1 s fundamental frequency values from speaker audio channel within turn (all 1 s intervals from turn start to end)
- *Loudness*: Average of speaker 1 s log energy values within turn
- *Loudness variation*: Standard deviation of 1 s log energy values within turn
- *Intensity*: Average of 1 s model-predicted intensity values within turn

S.1.1. Visual Features

All visual features were captured at 1 s intervals. Three classes of visual features were extracted: head movement, gaze, and facial emotion. Each visual feature was computed for both listener and speaker. To capture the objective visual signals of head nodding/shaking and gaze, we developed our own algorithmic detectors. Using facial recognition software in the Dlib C++ library, we computed a set of facial landmarks (83). Our head nodding/shaking detector then employed a manually tuned, rule-based approach that evaluated whether facial landmarks moved at least 10% of the total detected face size and crossed their starting position at least twice within two seconds. When this occurred along the vertical camera axis, we recorded a “nod,” generally taken as a nonverbal signal of “yes.” If it occurred along the horizontal axis, it was recorded as a “shake,” typically indicating “no.” Nods and shakes were aggregated to the turn level by

computing the maximum across all 1 s intervals in the turn, indicating whether any nodding or shaking occurred. Second, to measure whether participants were gazing at the screen, we used eye landmarks to compute the proportion of pixels within the eye regions that were white. If this proportion fell between 0.22 to 0.45, we estimated gaze to be directed at the screen. We caution that this variable appears to be noisy and has not been tested for accuracy. Finally, to obtain a proxy for facial emotion, we used FastAI (109) to train an emotion recognition model on the AffectNet corpus of facial expression images (84). AffectNet categorizes facial expressions into eight emotional groups; given the low estimated incidence of facial emotions other than happiness (and neutrality), we extracted only the predicted probability of happiness.

- *Listener/speaker nodding yes*: Vertical movement of facial landmarks exceeding a manually tuned threshold
- *Listener/speaker shaking no*: Horizontal movement of facial landmarks exceeding a manually tuned threshold
- *Listener/speaker gazing on screen*: White pixel proportion within eye region exceeding a manually tuned threshold
- *Listener/speaker facial happiness*: Predicted probability obtained from AffectNet-trained neural network

S.2. Statistical Methods

In this section, we describe our procedure for assessing whether groups of participants diverge in their conversational behavior. The same procedures are employed both for analyzing speech patterns by (a) partner-rated conversationalist score and (b) partner identity (see Manuscript Results Section 5 and Supplement Section B).

Our primary analysis of conversationalist score compares the outermost quartiles; additional results are given in the Supplement for all four quartiles.

Below, we first describe how we tested the null hypothesis that a conversational feature Y, such as loudness, was distributed equally among the K groups—or that $f(y \mid X=x_k) = f(y \mid X=x_{k'})$ for all $k, k' \in \{1, \ldots, K\}$. In other words, we evaluate whether every group k uses highly intense speech at the same rate as every other group k'. This approach utilized the full distribution of each feature and as such was well suited to capturing nonlinearities often observed in conversational data. However, a key limitation was that it produced *p* values as a test statistic and must be interpreted primarily by visually comparing feature distributions. To report numerical differences, the following subsection describes how we analyzed differences in means, $E[Y \mid k] - E[Y \mid k']$, and produced confidence intervals. Finally, we describe our procedure for accounting for multiple significance testing.

S.2.1. Assessing Differences in Distributions

Two key statistical challenges arose in these analyses. First, conversational features exhibit clustered dependence within a participant-conversation unit and across participants within a conversation. For example, idiosyncrasies in microphone positioning might cause speech from one participant to sound louder, or background noise might cause both participants to speak more loudly. Second, the number of observations (turns) within a cluster (conversations) can be influenced by the explanatory variables of interest (e.g., conversationalist score).

To test for differences in conversational patterns while accommodating these statistical issues, we first discretized each feature into deciles (i.e., quietest 10% of turns, turns between the

10^{th} and 20^{th} percentile on loudness, etc.) that captured much of the variation in how participants engaged with each other. That is, we represented each turn-level value, Y_i, with a one-hot encoding of the form $Y^*_i = [1, 0, 0, 0, 0, 0, 0, 0, 0, 0]$, here indicating a value of Y_i for i=1, (i.e., in the lowest decile). We then conducted a test of equal category proportions using an asymptotic multivariate Gaussian approximation for the multinomial distribution. Specifically, we constructed a matrix of turn-level discretized feature values, $\mathbf{Y}^* = [Y^*_{1,\setminus 1}{}^T, \ \ldots, Y^*_{N,\setminus 1}{}^T]^T$, where $Y^*_{i,\setminus 1} = [Y^*_{i,2}, \ \ldots, Y^*_{i,K}]$ represented the one-hot encoding of Y_i with the lowest reference category omitted (as category proportions sum to unity). Each turn was weighted to ensure that the total weight of each speaker in each conversation was equal, i.e., longer conversations received smaller weight on each turn. We then conducted a weighted multiple outcome regression of $\mathbf{Y}^*$ on group indicator variables $\mathbf{X}$, with $\mathbf{X} = [X_1, \ \ldots, X_N]$ and X_i as a k-dimensional one-hot vector in which a positive entry in the k^{th} position indicating the turn belonged to a participant with membership in group k. This produced a 9-dimensional vector of coefficient estimates for each group's categorical proportions, $\boldsymbol{\beta}_k = [\beta_{k,2}, \ \ldots, \beta_{k,10}]^T$, of turn features for that group (recalling that the omitted category proportion, $\beta_{k,1}$, sums to unity). Finally, we conducted an F test for the linear hypotheses of equality that coefficient vectors between each pair of groups—i.e., $\boldsymbol{\beta}_k = \boldsymbol{\beta}_{k'}$ for all k, k'—using an estimated variance-covariance matrix clustered at the conversation level. This approach had the advantage of easily accommodating conversation-level clustering and turn-level weights; it carried the disadvantage of requiring discretization of continuous features, discarding information, and resulting in some loss of statistical power. Alternative approaches based on clustered rank-sum tests (e.g., 110), which do not discretize the data, offer greater statistical power but are computationally infeasible in large datasets like the one studied here.

S.2.2. Assessing Differences in Winsorized Means

While tests of distributional equality are well-suited for assessing nonlinear differences in speech patterns, the *p* values they produce do not provide insight about precisely where and how those differences arise. For this reason, we provide distributional plots that convey, for example, how bad conversationalists are more likely to speak in a moderately loud voice, whereas good conversationalists are more polarized between quiet and loud speech.

To aid interpretation, we also computed differences in central tendency, which were straightforward to summarize and facilitated the reporting of confidence regions. This was complicated by the fact that automated processing occasionally resulted in outliers that strongly distorted simple averages. For example, in some cases, slight errors in the start/stop timestamps of short turns produced outlying values speech-speed estimates due to division by near-zero values. Similarly, audio artifacts occasionally arose from non-speech events such as laptop movement, producing outlying values that commanded disproportionate leverage in subsequent analyses. To address these issues, we employed Winsorization, a technique commonly used in analyses of audio data, for all unbounded variables (111, 112). Winsorizing at the (arbitrarily determined) 95% level replaced extreme values outside the 2.5^{th} and 97.5^{th} percentiles with the values of the boundary percentiles themselves. Finally, we conducted linear regressions of the resulting trimmed features on group indicators, obtaining estimated differences in Winsorized means. As in our distributional tests, we clustered standard errors at the conversation level and weighted turns to ensure that each speaker-conversation contributed equally to our estimates.

S.2.3. Multiple Testing Corrections

Our first set of mid-level analyses compared the best- and worst-rated conversationalists on 20 textual, auditory, and visual measures These tests resulted in 60 robustness analyses, as we repeated the same comparisons among subsets of female and male participants, as well as adjusting for turn duration. To control the false-discovery rate at conventional levels, we applied the multiple-testing correction of Benjamini & Hochberg (113) within each study. All reported *p* values were inflated by a corrective factor, ensuring they can be interpreted as usual (e.g., with reference to a 0.05 significance level) rather than utilizing a modified significance threshold.

S.3. Results from Descriptive Analysis of Patterns by Partner-rated Conversationalist Score

In this section, we report complete results and robustness tests from our study of high- and low-skilled conversationalists. Section S.3.1 presents a comprehensive set of results from our main analysis, including additional transcript-based, auditory, and visual features not reported in the main text as well as additional subgroups of "middling" conversationalists. Section S.3.2 contains results on female and male participants alone, allowing an assessment of gender heterogeneity of results. Section S.3.3. reports results after controlling for turn duration, allowing for an assessment of whether differences in non-duration conversational patterns such as speech speed or loudness may be in part driven by differences in turn duration. Finally, Section S.3.4. demonstrates that semantic similarity results are robust to the choice of a widely used alternative embedding model, RoBERTa, in place of the MPNet-based results presented in the main analyses.

S.3.1. Complete Results from Main Analysis

In this section, we present comprehensive findings from our study of how highly skilled conversationalists (as rated by their partners) differ from their low-skilled counterparts. Results

proceed as follows. Figures S.1, S.2, and S.3 respectively provide results on all transcript-based, auditory, and visual features; for completeness, we also report the estimated behavior of middling conversationalists (i.e., groups rated in the 25-50th percentile and 50-75th percentile) in addition to results on bad and good conversationalists (0-25th percentile and 75-100th percentile). Table S.3 provides a summary table assessing the statistical significance of differences in distributions between bad and good conversationalists, using *p* values corrected for multiple testing.

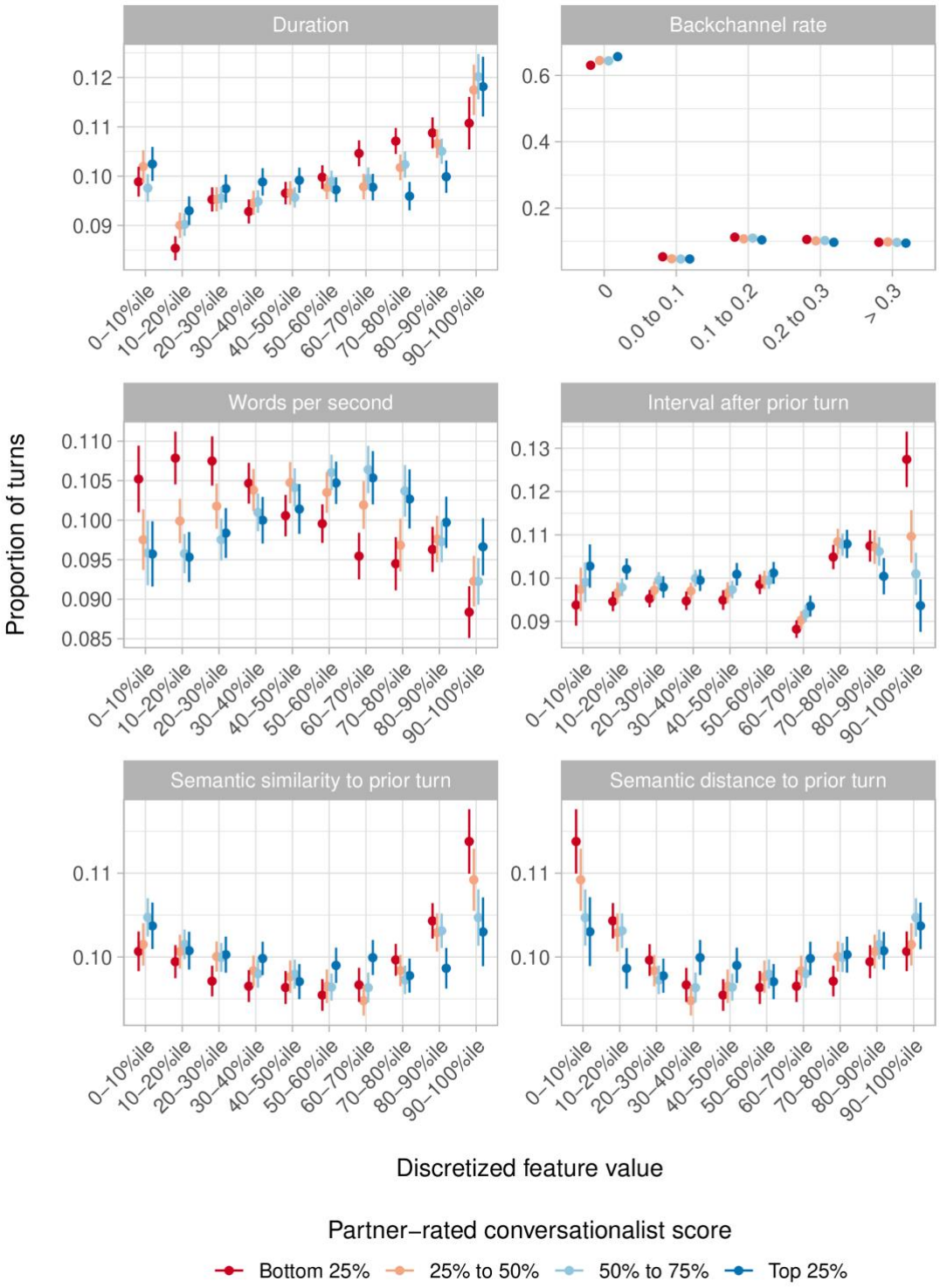

Fig. S.1. Behavior of good, middling, and bad conversationalists on transcript-based features. Each panel depicts the engagement patterns of good conversationalists (top 25% of partner-rated conversationalist score, depicted in blue) and bad conversationalists (bottom 25%, red) on a turn-level characteristic, expanding upon Figure 8 in the main text with additional panels. For completeness,

the plot also depicts middling conversationalists who are above the median (50–75th percentile, light blue) and below the median (25–50th percentile, light red). Horizontal axes denote categories of turn-level characteristics, defined in terms of feature deciles. The vertical position of each point indicates the average proportion of turns in a category for each group of conversationalists.

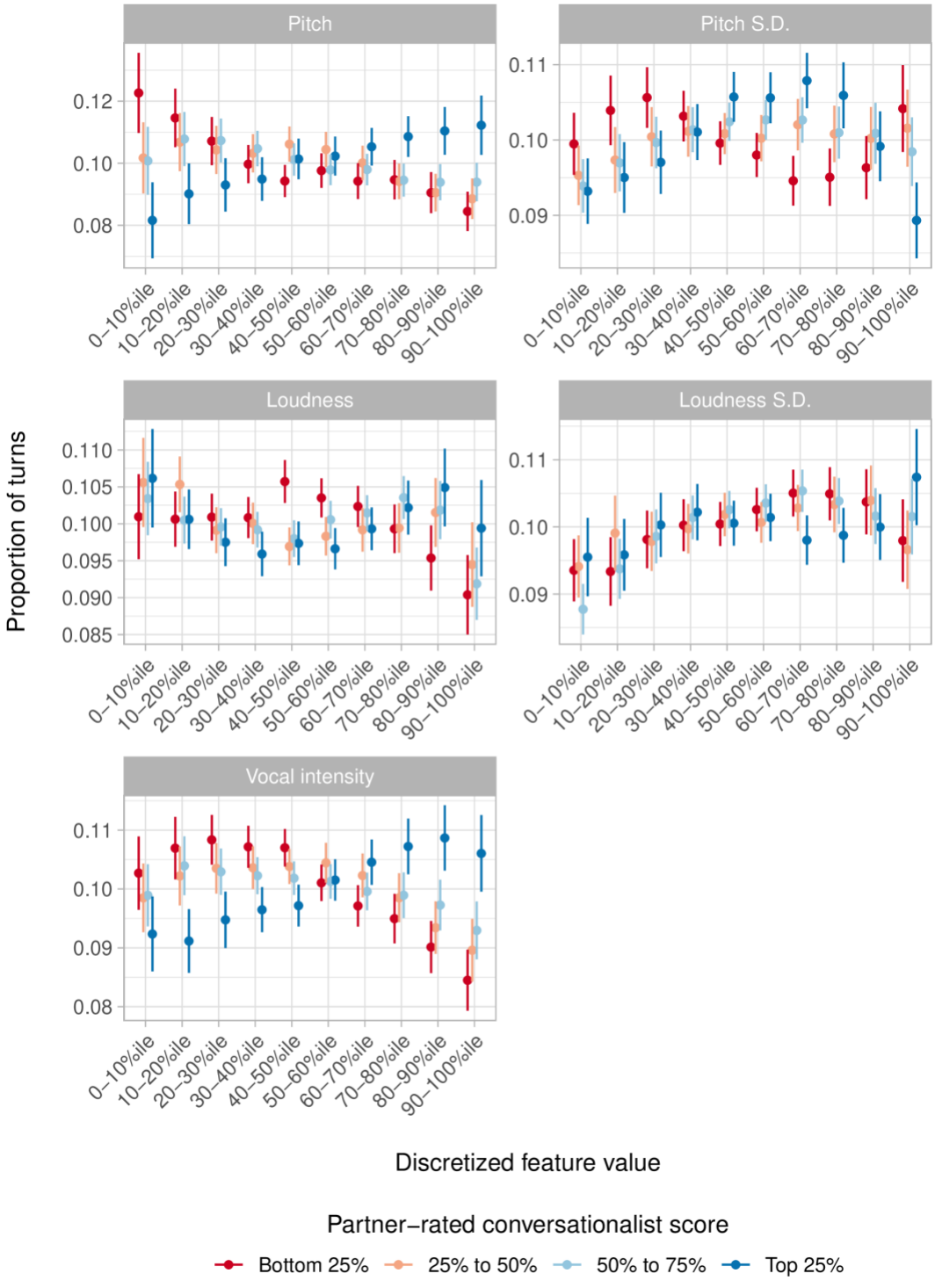

Fig. S.2. Behavior of good, middling, and bad conversationalists on auditory-based features. Each panel depicts the engagement patterns of good conversationalists (top 25% of partner-rated conversationalist score, depicted in blue) and bad conversationalists (bottom 25%, red) on a turn-level characteristic, expanding upon Figure 8 in the main text with additional panels. For completeness, the plot also depicts middling conversationalists who are above the median (50–75th percentile, light blue) and below the median (25–50th percentile, light red). Horizontal axes denote categories of turn-level characteristics, defined in terms of feature deciles. The vertical position of each point indicates the average proportion of turns in a category for each group of conversationalists.

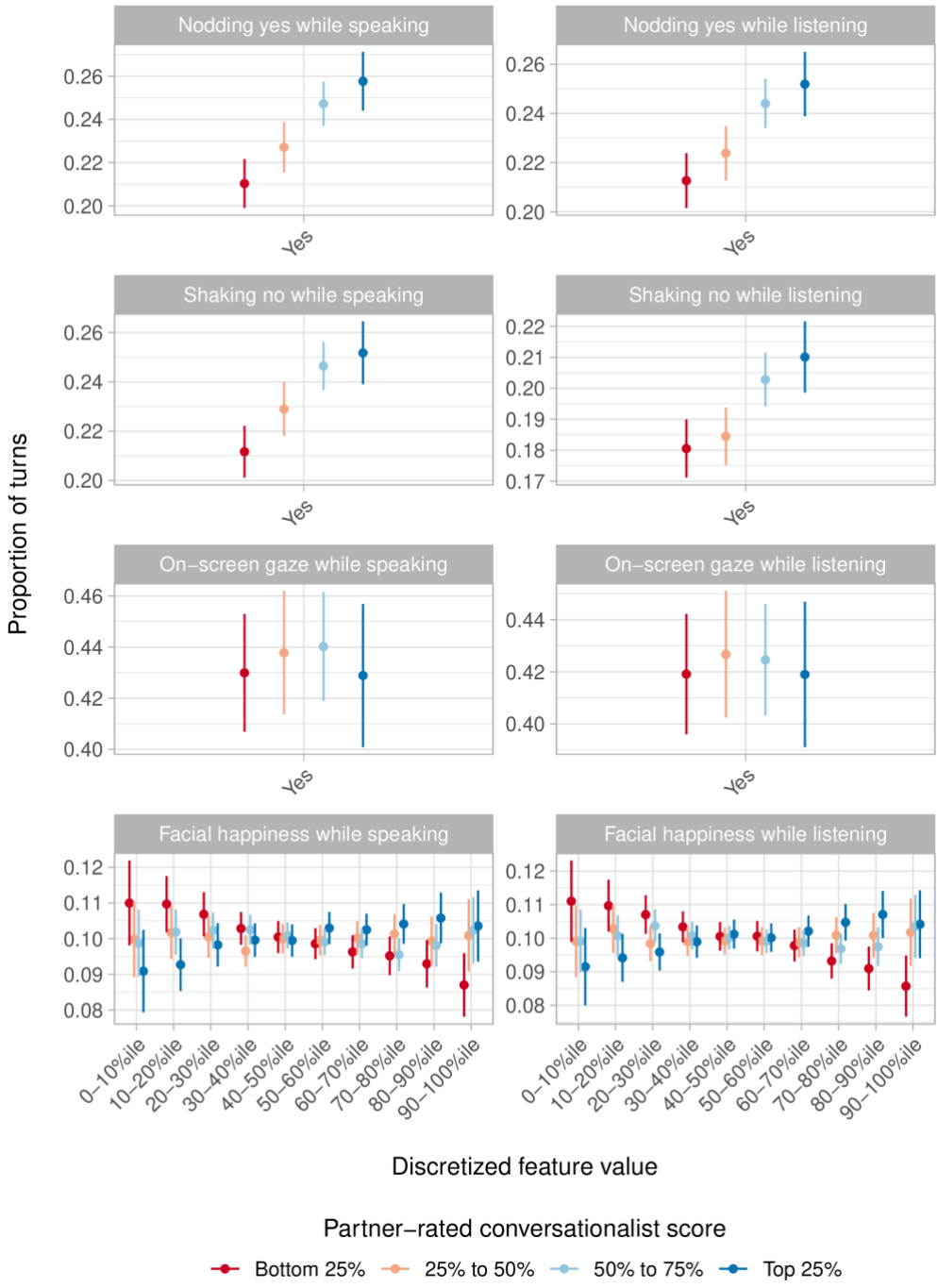

Fig. S.3. Behavior of good, middling, and bad conversationalists on visual-based features. Each panel depicts the engagement patterns of good conversationalists (top 25% of partner-rated conversationalist score, depicted in blue) and bad conversationalists (bottom 25%, red) on a turn-level characteristic, expanding upon Figure 8 in the main text with additional panels. For completeness, the plot also depicts middling conversationalists who are above the median (50–75th percentile, light blue) and below the median (25–50th percentile, light red). Horizontal axes denote categories of turn-level characteristics, defined in terms of feature deciles. The vertical position of each point indicates the average proportion of turns in a category for each group of conversationalists.

Table S.3. Statistical significance of differences in behavior between good and bad conversationalists (main results). Each row assesses differences between good conversationalists (top 25% of partner-rated conversationalist score) and bad conversationalists (bottom 25%) on a turn-level conversational feature. Separate *p* values are reported for tests of distributional equality and for tests of mean equality. The table reports only main analyses (all participants, no adjustment for turn duration), but multiple-testing adjustment accounts for robustness tests reported elsewhere (analyses restricted to male and female participants, as well as adjusting for turn duration).

	Feature	Diff.	95% CI	p_{adj} (mean)	p_{adj} (distr.)
Transcript	Interval after prior turn	-0.0523	[-0.0680, -0.0366]	<0.001	<0.001
Transcript	Duration	0.0959	[-0.1519, 0.3438]	0.608	<0.001
Transcript	Words per second	0.0996	[0.0595, 0.1396]	<0.001	<0.001
Transcript	Backchannel rate	-0.0046	[-0.0079, -0.0013]	0.016	<0.001
Transcript	Cosine similarity to prior	-0.0053	[-0.0081, -0.0025]	0.001	0.001
Transcript	Euclidean dist. to prior	0.004	[0.0014, 0.0067]	0.009	0.001
Auditory	Vocal intensity	0.0131	[0.0072, 0.0190]	<0.001	<0.001
Auditory	Pitch	11.18	[6.84, 15.53]	<0.001	<0.001
Auditory	Loudness	0.1189	[-0.4229, 0.6608]	0.809	0.016
Auditory	Pitch S.D.	-0.2609	[-1.2030, 0.6812]	0.744	<0.001
Auditory	Loudness S.D.	0.0555	[-0.3872, 0.4982]	0.835	0.025
Visual	Facial happiness (listening)	0.0354	[0.0150, 0.0557]	0.003	0.045
Visual	On-screen gaze (listening)	-0.0071	[-0.0430, 0.0288]	0.81	0.994
Visual	Nodding yes (listening)	0.0401	[0.0215, 0.0588]	<0.001	<0.001
Visual	Shaking no (listening)	0.0298	[0.0134, 0.0461]	0.002	<0.001
Visual	Facial happiness (speaking)	0.0318	[0.0118, 0.0518]	0.006	0.255
Visual	On-screen gaze (speaking)	-0.0059	[-0.0417, 0.0300]	0.821	0.981
Visual	Nodding yes (speaking)	0.0495	[0.0302, 0.0687]	<0.001	<0.001
Visual	Shaking no (speaking)	0.041	[0.0230, 0.0591]	<0.001	<0.001

S.3.2. Gender-disaggregated Results

In this section, we present additional findings from robustness tests that subset female and male respondents before comparing high- and low-skilled conversationalists.

Fig. S.4. Behavior of good and bad conversationalists (female participants). Each panel depicts the engagement patterns of good conversationalists (top 25% of partner-rated conversationalist score, depicted in blue) and bad conversationalists (bottom 25%, red) on a turn-level characteristic. Horizontal axes denote categories of turn-level characteristics, defined in terms of feature deciles. The vertical position of each point indicates the average proportion of turns in a category for good or bad conversationalists.

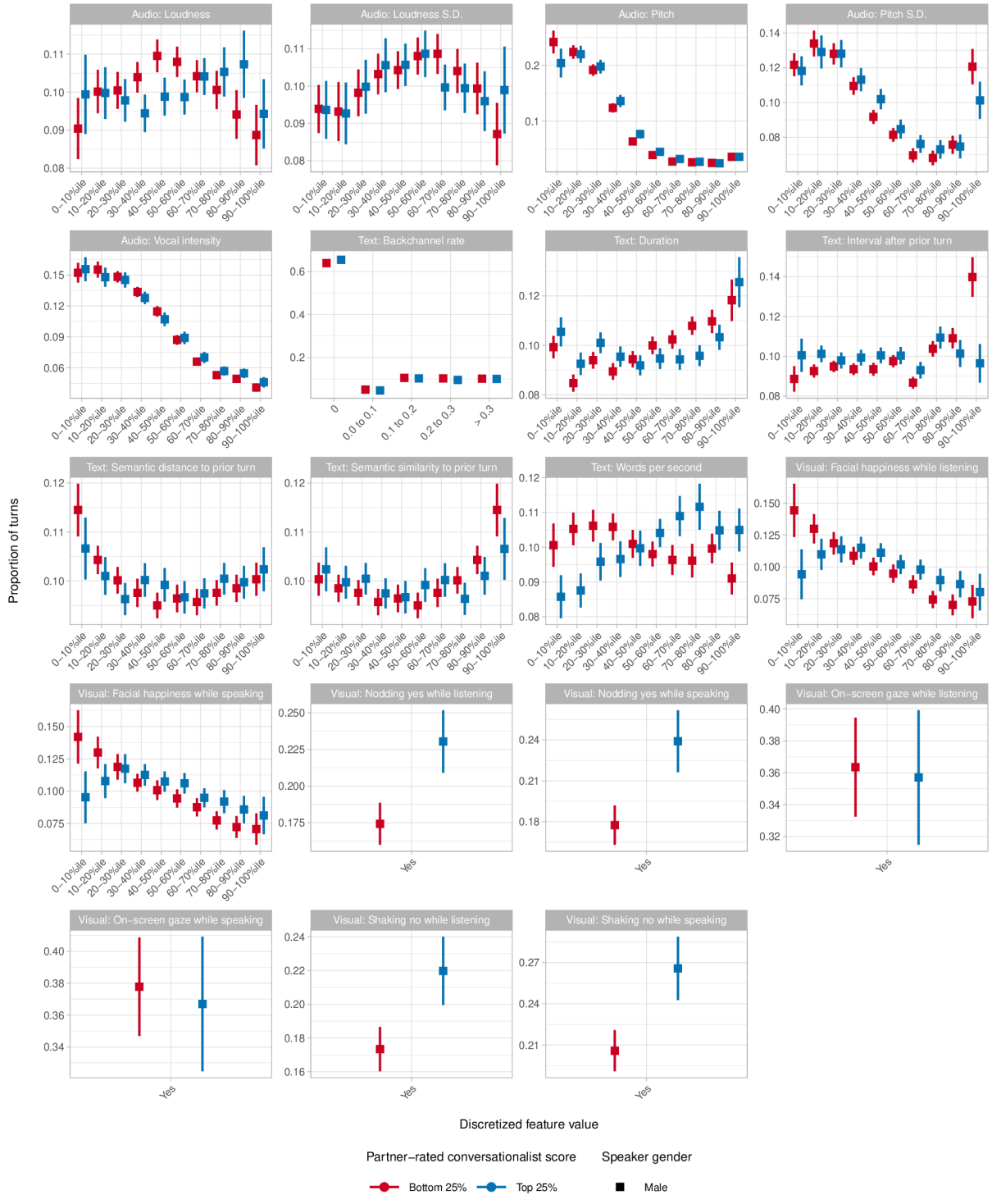

Fig. S.5. Behavior of good and bad conversationalists (male participants). Each panel depicts the engagement patterns of good conversationalists (top 25% of partner-rated conversationalist score, depicted in blue) and bad conversationalists (bottom 25%, red) on a turn-level characteristic. Horizontal axes denote categories of turn-level characteristics, defined in terms of feature deciles. The vertical

position of each point indicates the average proportion of turns in a category for good or bad conversationalists.

Table S.4. Statistical significance of differences in behavior between good and bad conversationalists (female participants). Each row assesses differences between good conversationalists (top 25% of partner-rated conversationalist score) and bad conversationalists (bottom 25%) on a turn-level conversational feature within a participant gender. Separate *p* values are reported for tests of distributional equality and for tests of mean equality. The table reports only analyses among female participants, but multiple-testing adjustment accounts for additional tests discussed elsewhere

	Feature	Diff.	95% CI	p_{adj} (mean)	p_{adj} (distr.)
Transcript	Interval after prior turn	-0.034	[-0.0550, -0.0139]	0.003	0.060
Transcript	Duration	0.270	[-0.0498, 0.5893]	0.164	<0.001
Transcript	Words per second	0.060	[0.0091, 0.1107]	0.044	0.212
Transcript	Backchannel rate	-0.005	[-0.0097, -0.0004]	0.065	<0.001
Transcript	Cosine similarity to prior	-0.006	[-0.0094, -0.0017]	0.013	0.038
Transcript	Euclidean dist. to prior	0.004	[0.0003, 0.0078]	0.065	0.038
Auditory	Vocal intensity	0.005	[0.0004, 0.0103]	0.065	0.069
Auditory	Pitch	4.730	[1.42, 8.05]	0.014	0.002
Auditory	Loudness	0.141	[-0.6036, 0.8858]	0.810	0.691
Auditory	Pitch S.D.	0.200	[-0.7544, 1.1543]	0.809	0.043
Auditory	Loudness S.D.	-0.062	[-0.6728, 0.5495]	0.858	0.735
Visual	Facial happiness (listening)	0.020	[-0.0067, 0.0475]	0.223	0.195
Visual	On-screen gaze (listening)	-0.006	[-0.0546, 0.0422]	0.835	0.710
Visual	Nodding yes (listening)	0.019	[-0.0069, 0.0442]	0.233	0.259
Visual	Shaking no (listening)	0.019	[-0.0024, 0.0405]	0.145	0.155
Visual	Facial happiness (speaking)	0.015	[-0.0115, 0.0419]	0.376	0.392
Visual	On-screen gaze (speaking)	-0.006	[-0.0543, 0.0421]	0.835	0.786
Visual	Nodding yes (speaking)	0.033	[0.0063, 0.0591]	0.035	0.043
Visual	Shaking no (speaking)	0.030	[0.0067, 0.0536]	0.029	0.038

Table S.5. Statistical significance of differences in behavior between good and bad conversationalists (male participants). Each row assesses differences between good conversationalists (top 25% of partner-rated conversationalist score) and bad conversationalists (bottom 25%) on a turn-level conversational feature within a participant gender. Separate *p* values are reported for tests of distributional equality and for tests of mean equality. The table reports only analyses among male participants, but multiple-testing adjustment accounts for additional tests discussed elsewhere.

	Feature	Diff.	95% CI	p_{adj} (mean)	p_{adj} (distr.)
Transcript	Interval after prior turn	-0.066	[-0.0914, -0.0410]	<0.001	<0.001
Transcript	Duration	0.070	[-0.3307, 0.4710]	0.818	0.015
Transcript	Words per second	0.157	[0.0919, 0.2220]	<0.001	<0.001
Transcript	Backchannel rate	-0.003	[-0.0081, 0.0021]	0.361	0.192
Transcript	Cosine similarity to prior	-0.004	[-0.0079, -0.0000]	0.091	0.324
Transcript	Euclidean dist. to prior	0.003	[-0.0005, 0.0067]	0.160	0.322
Auditory	Vocal intensity	0.002	[-0.0053, 0.0095]	0.744	0.193
Auditory	Pitch	2.849	[-0.9402, 6.6389]	0.223	0.255

Auditory	Loudness	0.017 [-0.8124, 0.8456]	0.969	0.021
Auditory	Pitch S.D.	-1.279 [-3.04, 0.48]	0.233	0.030
Auditory	Loudness S.D.	0.165 [-0.5010, 0.8310]	0.777	0.141
Visual	Facial happiness (listening)	0.038 [0.0079, 0.0687]	0.032	0.053
Visual	On-screen gaze (listening)	-0.022 [-0.0746, 0.0302]	0.565	0.855
Visual	Nodding yes (listening)	0.058 [0.0293, 0.0864]	<0.001	<0.001
Visual	Shaking no (listening)	0.046 [0.0192, 0.0737]	0.003	<0.001
Visual	Facial happiness (speaking)	0.037 [0.0068, 0.0665]	0.035	0.060
Visual	On-screen gaze (speaking)	-0.019 [-0.0715, 0.0337]	0.638	0.744
Visual	Nodding yes (speaking)	0.064 [0.0345, 0.0942]	<0.001	<0.001
Visual	Shaking no (speaking)	0.061 [0.0312, 0.0910]	<0.001	<0.001

S.3.3. Duration-adjusted Results

In this section, we present additional findings from robustness tests that controlled for turn duration in comparing high- and low-skilled conversationalists. To do so, we included demeaned turn duration as a linear predictor in the regressions described in Appendix S.2.1. This allowed for turn proportions in each category (lowest decile of a feature, second-lowest decile, etc.) to increase or decrease linearly as a function of duration. For example, the model allowed for a slight reduction in extremely loud speech (as measured by average decibels over the turn) for each additional second that the turn continued; this accounted for the possibility that, for instance, loud speech was difficult to sustain for long periods. At the same time, the model allowed for differently sized increases or decreases of other loudness categories (e.g., moderately quiet speech) for each additional second of turn duration. Figure S.6 depicts the predicted engagement patterns of good and bad conversationalists, holding turn duration fixed at the average value across the corpus. Table S.5 summarizes the statistical significance of duration-adjusted differences between high- and low-skilled conversationalists.

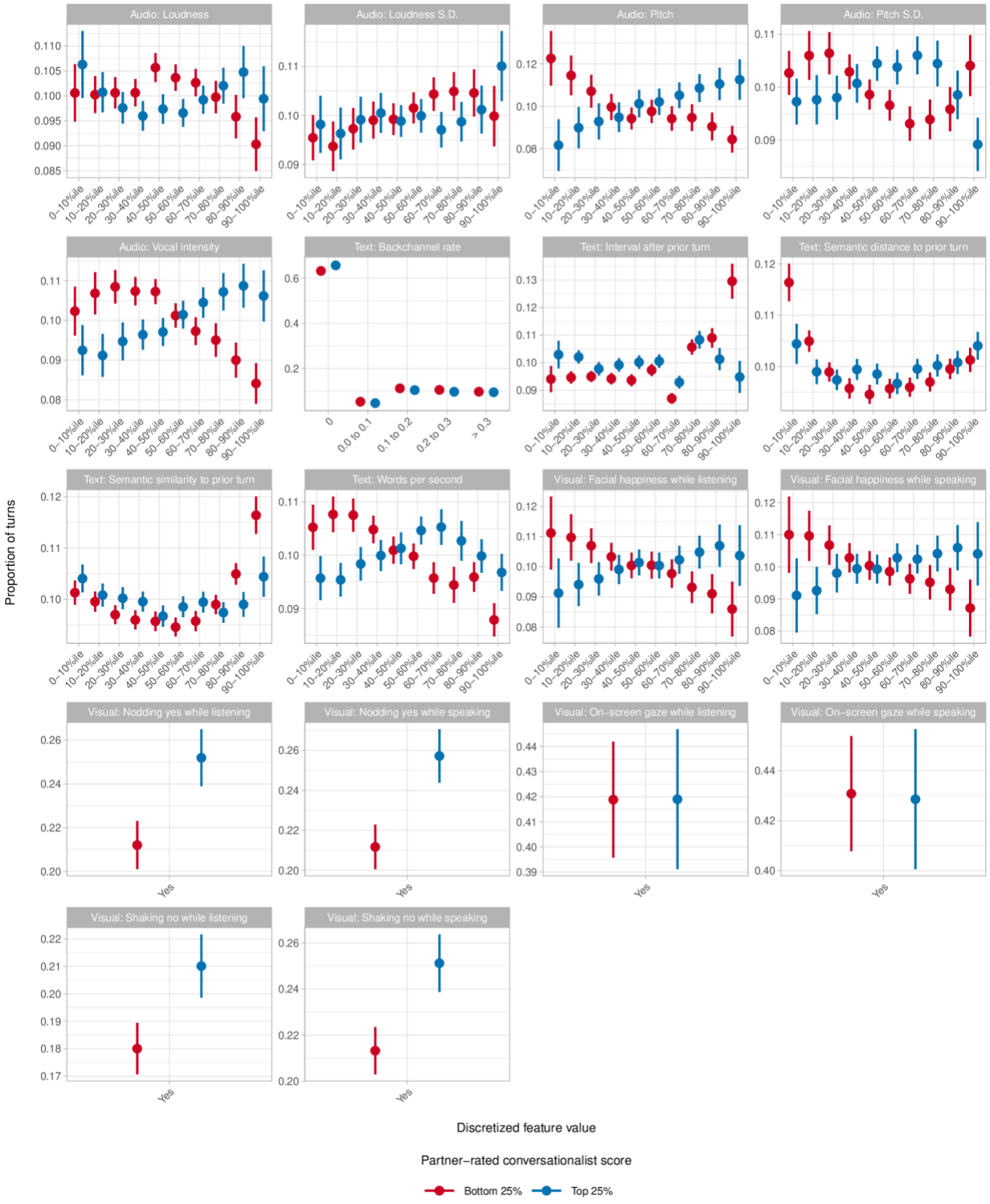

Fig. S.6. Behavior of good and bad conversationalists (duration-adjusted results). Each panel depicts the engagement patterns of good conversationalists (top 25% of partner-rated conversationalist score, depicted in blue) and bad conversationalists (bottom 25%, red) on a turn-level characteristic. Horizontal axes denote categories of turn-level characteristics, defined in terms of feature deciles. The vertical position of each point indicates the average proportion of turns in a category for good or bad conversationalists.

S.3.4. Robustness of Semantic Similarity Results

Finally, we demonstrate that results on novelty and semantic similarity were not simply idiosyncratic artifacts of the particular embedding model we used to reduce turn transcripts to a quantitative representation. While the MPNet embedding model used in our main analyses was selected on the basis of achieving the highest average performance across a number of domains, RoBERTa embeddings (74) are a widely used alternative. In broad strokes, Figure S.7 replicates the overall pattern of our findings: bad conversationalists had higher average similarity to the previous turn, indicating more repetitive, less novel, responses. However, we find that these differences did not manifest in a consistent manner across embedding models. MPNet results suggested that bad conversationalists were differentiated by a large number of extremely high-similarity statements that were near duplicates of prior turns. In contrast, RoBERTa results indicated that bad conversationalists had fewer high-novelty (low-similarity) statements. While both approaches suggested that poor conversationalists' contributions were more mundane, the practical implications of the different models' results present substantially different interpretations. Resolving the apparent divergence between these approaches may constitute an important area for future work.

We also found that certain methods to measure semantic distance appeared to depend heavily upon the number of words in current and previous turns, although importantly, our primary specification and results were robust to residualizing on word count. Finally, most embedding models are trained to interpret the intricacies of language based upon corpora of written documents, and not transcripts of spoken communication. As such, the rules that these models learn about language, and the subsequent numerical representations they produce, may not be fully suitable for a corpus such as ours, which is composed of entirely natural

conversations. Application of domain-transfer techniques may help address this gap, although success may be difficult to evaluate without extensive human annotation. For these reasons, we urge readers to exercise caution in interpreting the association between semantic novelty and conversationalist quality. Nevertheless, we consider it encouraging that previous work has found that pre-trained models can, in fact, achieve near-human performance across a range of domains and hyperparameter choices (Rodriguez & Spirling, 2002).

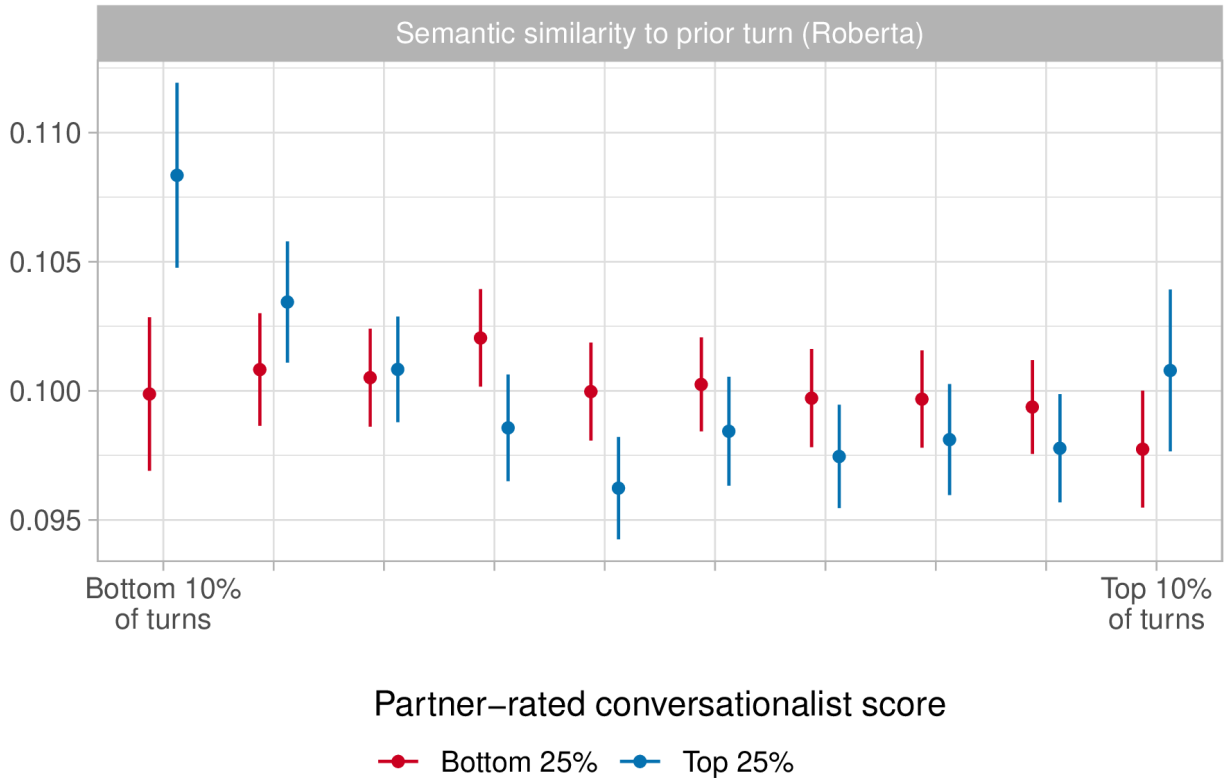

Fig. S.7. Behavior of good and bad conversationalists on RoBERTa embedding similarity. Cosine similarity of current turn to partner's prior turn for good conversationalists (top 25% of partner-rated conversationalist score, depicted in blue) and bad conversationalists (bottom 25%, red). Horizontal axes denote categories of semantic similarity, defined in terms of deciles across the corpus. The vertical position of each point indicates the average proportion of turns in a category for good or bad conversationalists.

S.3.5. Vocal Intensity Varies with Personality

In the main text of the manuscript, we compared the vocal intensity distributions of good and bad conversationalists, which differed significantly (i.e., the null of equal distributions was rejected at $p_{adj}<0.001$). People rated as good conversationalists spoke with greater vocal intensity than bad conversationalists. During revision, we were encouraged to explore further sources of variability in these effects such as people's personality traits. To do so, we examined how our good-bad conversationalist effects varied with people's scores on the Big 5 personality traits (i.e., whether people are above or below the median score for Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness).

For example, consider the neuroticism panel in Figure S.8, which shows that neuroticism moderates the vocal intensity of both good and bad conversationalists. For bad conversationalists, those higher in neuroticism spoke with greater vocal intensity compared to those lower in neuroticism (diff in average intensity score = 0.014, $p < 0.001$). A similar effect was observed for good conversationalists – again those higher in neuroticism spoke with greater vocal intensity compared to those lower in neuroticism (diff in average intensity score = 0.013, $p < 0.01$). In other words, being high in neuroticism appears to be associated with speaking with more vocal intensity during conversation, but this relationship between neuroticism and vocal intensity did not depend on whether someone was a good or bad conversationalist.

Neuroticism is a personality trait that appears to affect the behavior of good *and* bad conversationalists similarly. But one can also imagine a case where personality affects good and bad conversationalists differently. Consider the agreeableness panel of Figure 3. Here agreeableness only moderates the vocal intensity of good conversationalists. In other words, people's level of agreeableness does not affect their vocal intensity when they are bad conversationalists (diff in average intensity score = 0.002, $p = 0.64$). But agreeableness does

moderate the vocal intensity of good conversationalists, such that good conversationalists who are also high in agreeableness speak with more vocal intensity compared to good conversationalists who are low in agreeableness (diff in average intensity score = 0.010, $p < 0.05$). In our manuscript, we saw that good conversationalists speak with more vocal intensity compared to bad conversationalists, and here we see that the agreeableness further moderates this effect for good conversationalists specifically.

In sum, neuroticism and agreeableness are two personality traits that intuitively should be related to emotional intensity during conversation. We see evidence for this in our data. For neuroticism, higher vocal intensity may be related to the stress or anxiety that people feel in their initial interactions with strangers, and for agreeableness, perhaps such intensity is more associated with the positive emotions that agreeable people emphasize while being cooperative, polite, and friendly in first impression contexts. Furthermore, our analyses reveal how the relationship between personality and various behavioral features, such as vocal intensity, can sometimes remain stable across good and bad conversationalists (e.g., as we observed with neuroticism), and other times, how the effect of personality only emerges in conjunction with being a good conversationalist, but not a bad conversationalist (e.g., as we observed with agreeableness). Overall, we identified a number of behavioral patterns that distinguish good and bad conversationalists, and personality appears to moderate these findings in nuanced ways.

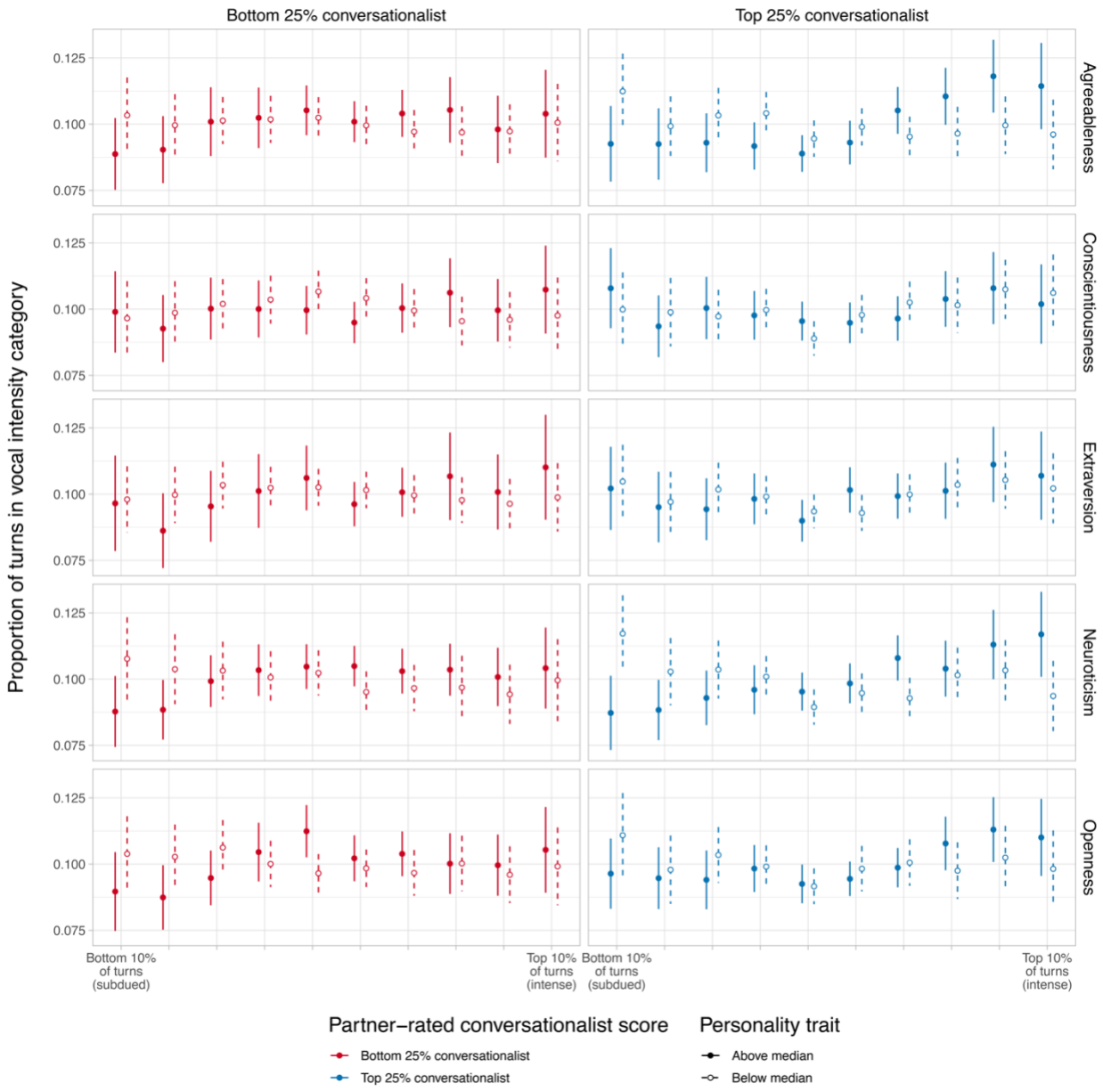

Fig. S.8. Each panel depicts the vocal intensity of good conversationalists (top 25% of partner-rated conversationalist score, depicted in blue) and bad conversationalists (bottom 25%, depicted in red). Results are further divided by people's scores on the Big 5 personality traits (i.e., whether people are above or below the median score for Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness). Horizontal axes denote turn-level feature deciles. The y-axis indicates the mean proportion of turns in a category for a good or bad conversationalist. Error bars represent 95% confidence intervals.

Section on Topical, Relational, and Demographic Diversity

In this section, we present results from a quasi-experimental analysis that examined how members of one identity group shift their conversational patterns when quasi-randomly assigned to partners of (1) their own group, or (2) partners of a differing group.

The same procedure assessing whether K groups of participants diverge in their conversational behavior was applied here; for a detailed discussion, see Supplement Section S.2. Here we first restrict analysis to participants from one demographic (e.g., older participants), then examine whether those participants engage in conversation differently when assigned to older, middling, or younger partners, i.e. among K=3 subgroups of older participants). Similarly, to correct for multiple testing, the Benjamini-Hochberg procedure described in Supplement Section S.2.3 was applied here to 340 tests about quasi-randomly assigned partner identity on participant behavior (involving nine subgroups of participants and 17 contrasts between in-group partners and various out-groups of partners, again repeated on 20 features).

All analyses compare *within* a group (e.g. subsetting to young participants), making contrasts within that subset based on the group of the assigned partner (e.g., comparing those assigned to old partners, as opposed to young partners). We emphasize that young and old participants differ in many ways, such as their education level and political attitudes. Our analysis does not seek to disentangle which specific attribute drives the difference in engagement patterns—that is, it does not claim that effects are due to the age gap alone, holding all other attributes fixed. Rather, it aims to approximate an ideal experiment in which a participant is randomly assigned to converse with a partner from group A or B, where A and B differ on some aspect of identity as well as the "bundle of sticks" (91) that are associated with or comprise that

identity. Moreover, it does not attempt to identify psychological mechanisms underlying the change in an individual's behavior, such as out-group animosity. Finally, we note that as discussed in the main text, differences in one participant's behavior can arise as a response to differing behavior by another participant. Throughout, reported *p* values are adjusted for the multiplicity of features analyzed and partner-group comparisons made.

Partner assignment is based on an algorithm that greedily matches pairs of participants that indicated their availability during the same time slot. Because the matching algorithm does not incorporate demographic information, whether a participant is assigned to an in-group or out-group partner is guaranteed to be ignorable conditional on availability. For purposes of analysis, we assume that it is ignorable when aggregating over availability blocks as well. To assess the plausibility of this design assumption, we conduct chi-squared tests to evaluate dependence in participant and partner identity—for example, whether older participants are more likely to be paired with other older participants, compared to a null model in which they are randomly assigned to partners of all ages. Chi-squared tests for dependence in age, gender, race, education, and political ideology pairings respectively produce *p* values of 0.53, 0.33, 0.62, 0.91, and 0.43. These results suggest that availability is at most weakly related to membership in an identity group. Moreover, within identity groups, we assess that availability is unlikely to correlate strongly with baseline conversational patterns.

In what follows, Sections S.4.1–3 respectively present results on quasi-randomly assigned partner age, gender, and race/ethnicity. We do not detect significant differences in conversational patterns by partner education (distributed 37%, 40%, and 23% respectively below, at, and above the level of a bachelor's degree) or political ideology (65% liberal, 20% neutral, 15% conservative), though we caution that the statistical power of political-ideology

results is limited by the relatively small proportion of conservative participants. These results are omitted to conserve space, but reported *p* values include adjustment for all analyses that were conducted. To aid interpretation, Table S.6 in Section S.4.4 summarizes differences in average feature values with 95% confidence intervals; multiple-testing-adjusted *p* values are reported for tests of differences in means.

S.4.1. Quasi-random Partner Age Results

To analyze age, we divide participants approximately into tertiles representing the youngest (aged 19–28; N=1,204), middle (29–38; N=1,013), or oldest age groups (39–66; N=1,039). Note that tertiles are slightly imbalanced due to rounding in reported age. Figures S.9, S.10, and S.11 respectively present analyses that subset to young, middle, and oldest participants, examining the distribution of their conversational features when paired with in- and out-group partners. For compactness, we plot only results that are statistically significant at the 0.05 level.

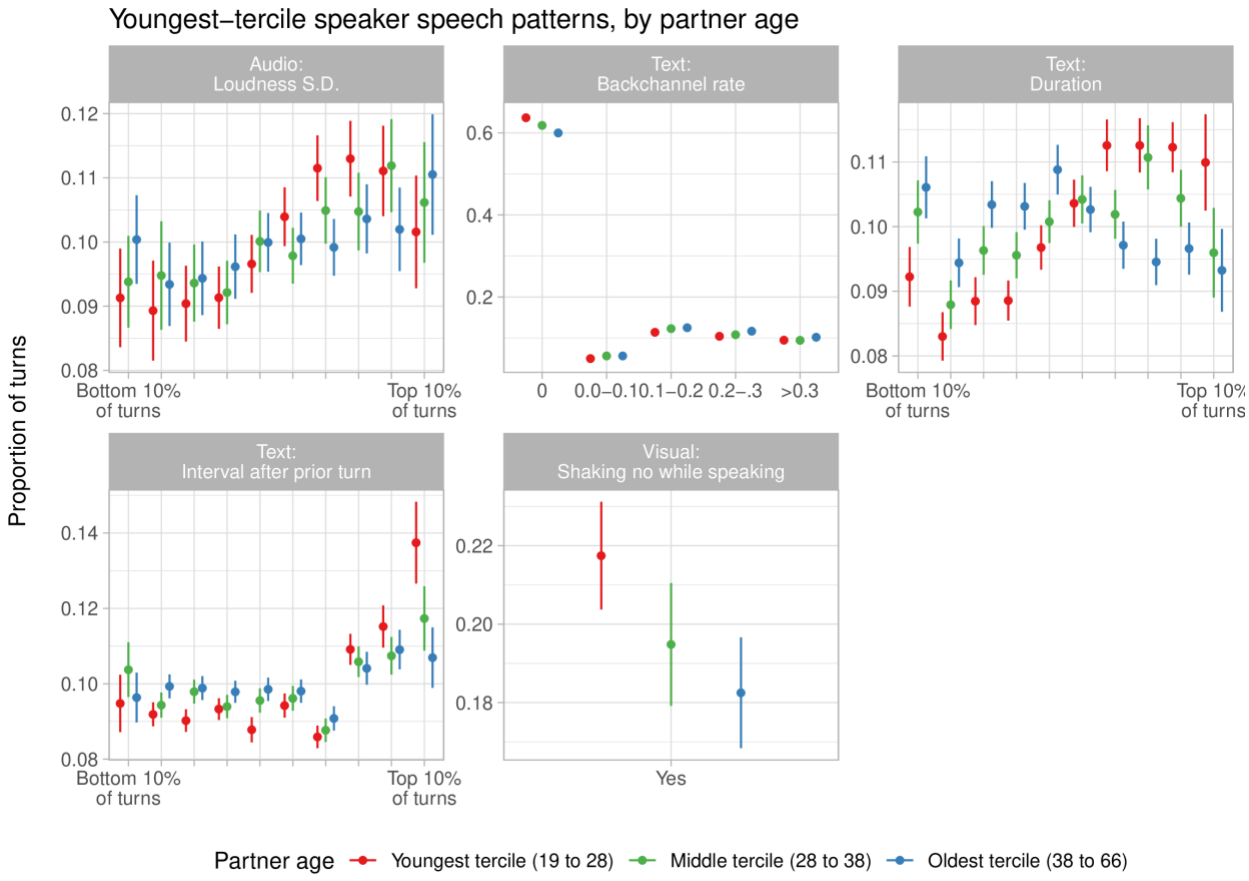

Fig. S.9. Behavior of youngest-tertile participants when assigned to young, middle, and old age-group partners. Each panel depicts the engagement patterns of young participants assigned to young (red), middle (green), or old (blue) age-group partners on a turn-level characteristic. Horizontal axes denote categories of turn-level characteristics, defined in terms of feature deciles. The vertical position of each point indicates the average proportion of turns in a category. Distributions are presented only for features in which the null hypothesis of distributional equality is rejected at the 0.05 level after multiple-testing adjustment.

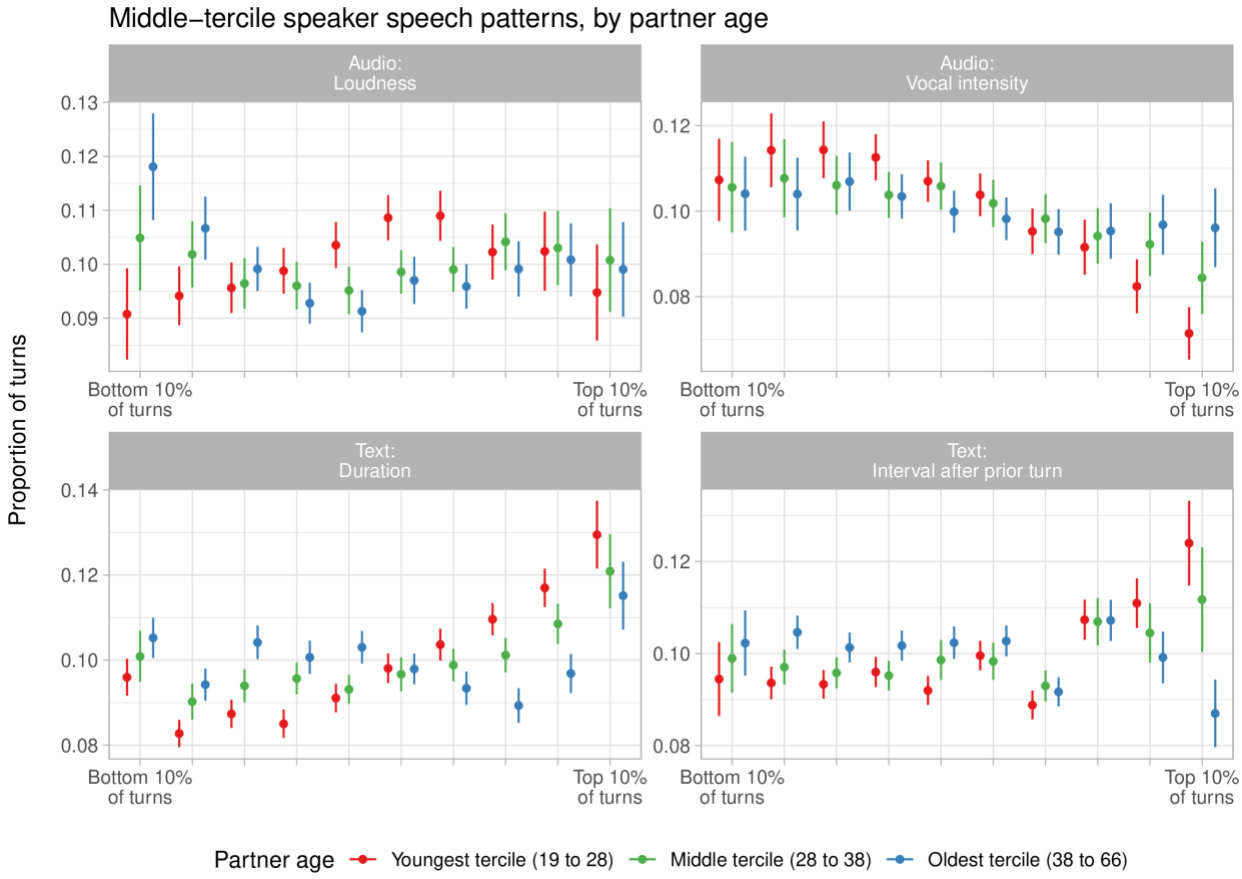

Fig. S.10. Behavior of middle age-group participants when assigned to young, middle, and old age-group partners. Each panel depicts the engagement patterns of middle age-group participants assigned to young (red), middle (green), or old (blue) age-group partners on a turn-level characteristic. Horizontal axes denote categories of turn-level characteristics, defined in terms of feature deciles. The vertical position of each point indicates the average proportion of turns in a category. Distributions are presented only for features in which the null hypothesis of distributional equality is rejected at the 0.05 level after multiple-testing adjustment.

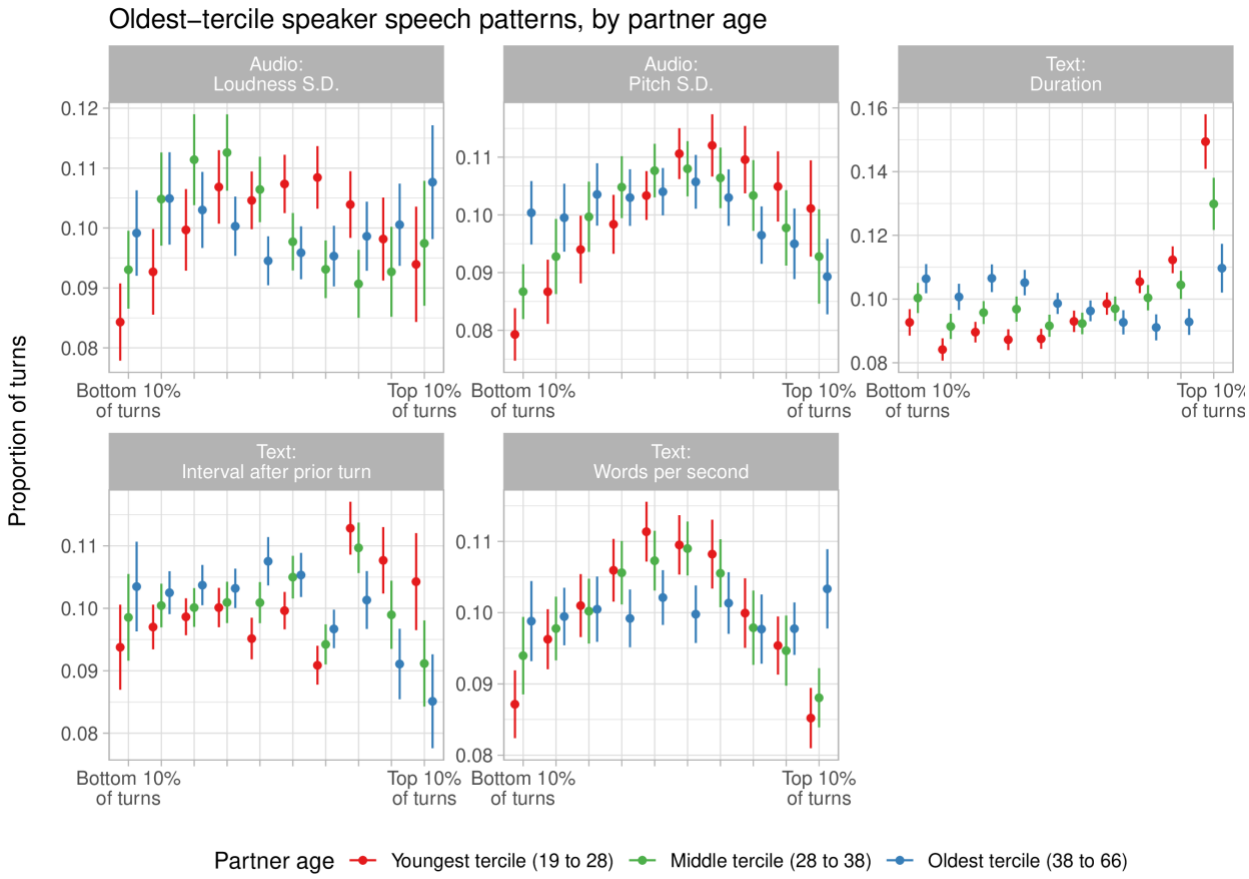

Fig. S.11. Behavior of oldest-tertile participants when assigned to young, middle, and old age-group partners. Each panel depicts the engagement patterns of old participants assigned to young (red), middle (green), or old (blue) age-group partners on a turn-level characteristic. Horizontal axes denote categories of turn-level characteristics, defined in terms of feature deciles. The vertical position of each point indicates the average proportion of turns in a category. Distributions are presented only for features in which the null hypothesis of distributional equality is rejected at the 0.05 level after multiple-testing adjustment.

S.4.2. Quasi-random Partner Gender Results

To analyze gender, we examine participants who self-describe as female (N=1,740) or male (N=1,463). Participants with other gender identities, as well as those who preferred not to answer, were not analyzed due to a lack of statistical power (N=109). Figures S.12 and S.13 respectively present analyses that subset to female and male participants, examining the distribution of their conversational features when paired with in- and out-group partners. For compactness, we plot only results that are statistically significant at the 0.05 level.

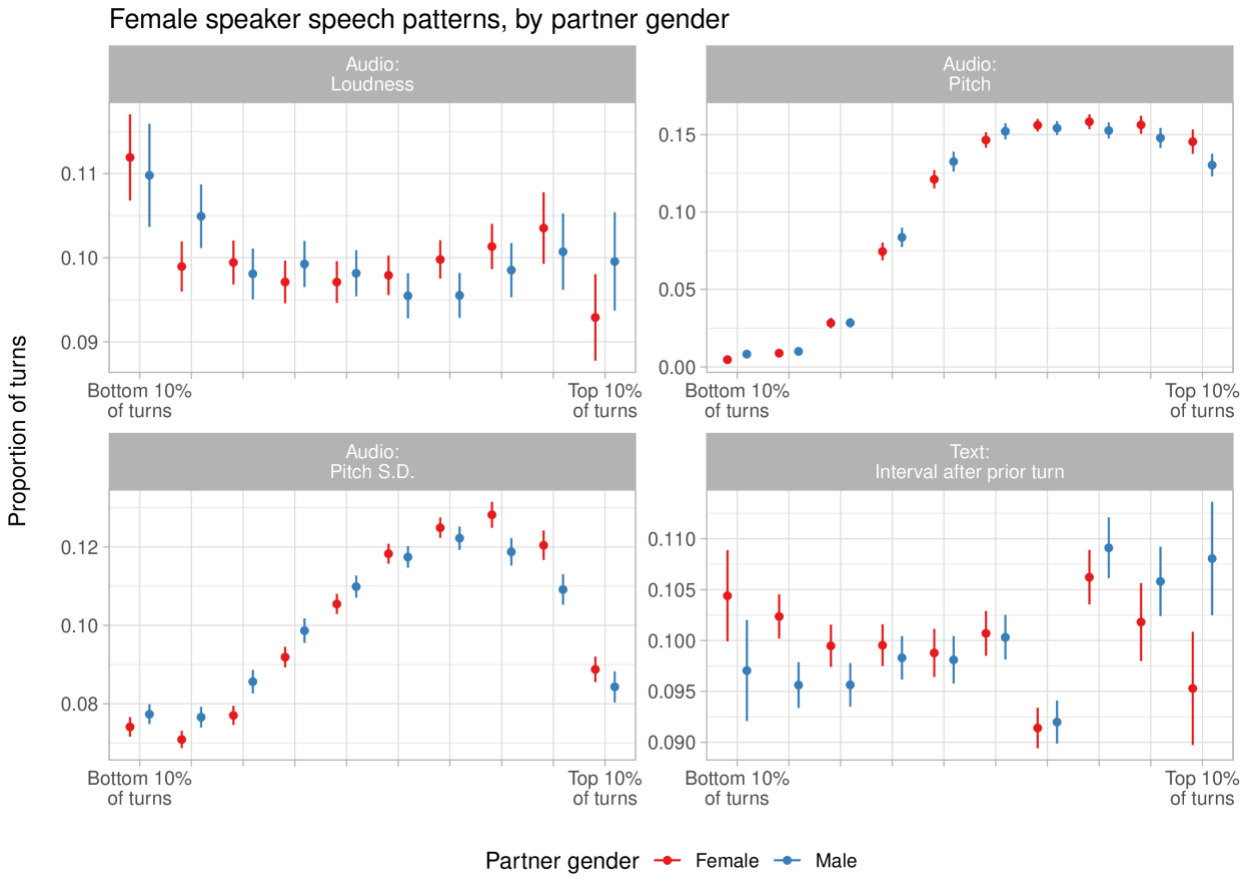

Fig. S.12. Behavior of female participants when assigned to female and male partners. Each panel depicts the engagement patterns of female participants assigned to female (red) or male (blue) partners on a turn-level characteristic. Horizontal axes denote categories of turn-level characteristics, defined in terms of feature deciles. The vertical position of each point indicates the average proportion of turns in a category. Distributions are presented only for features in which the null hypothesis of distributional equality is rejected at the 0.05 level after multiple-testing adjustment.

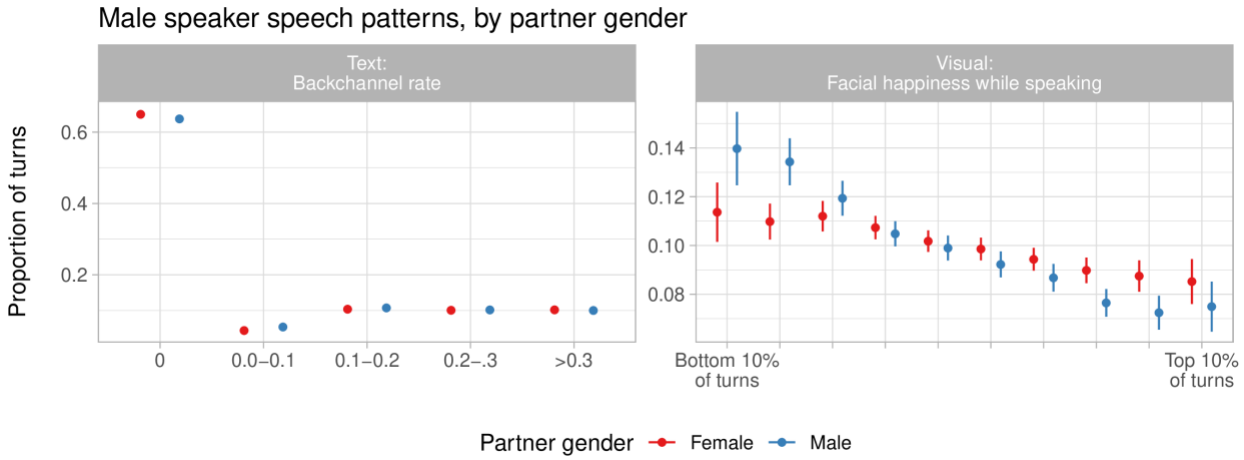

Fig. S.13. Behavior of male participants when assigned to female and male partners. Each panel depicts the engagement patterns of male participants assigned to female (red) or male (blue) partners on a turn-level characteristic. Horizontal axes denote categories of turn-level characteristics, defined in terms of feature deciles. The vertical position of each point indicates the average proportion of turns in a category. Distributions are presented only for features in which the null hypothesis of distributional equality is rejected at the 0.05 level after multiple-testing adjustment.

S.4.3. Quasi-random Partner Race/Ethnicity Results

To analyze race and ethnicity, we examine participants self-describing as Asian (N=485, 16%), Black (N=248, 8%), Hispanic (N=220, 7%), or White (N=2,110, 69%). These proportions roughly track the U.S. population (6% Asian, 13% Black, 19% Hispanic, and 60% White in 2021 Census data) but under-represent Black and Hispanic groups. Figure S.14 subsets to White participants and examines the distribution of their conversational features when paired with in- and out-group partners. Analyses of behavior by non-White groups is not feasible in this dataset due to the sparsity of minority-minority pairings. For compactness, we plot only results that are statistically significant at the 0.05 level

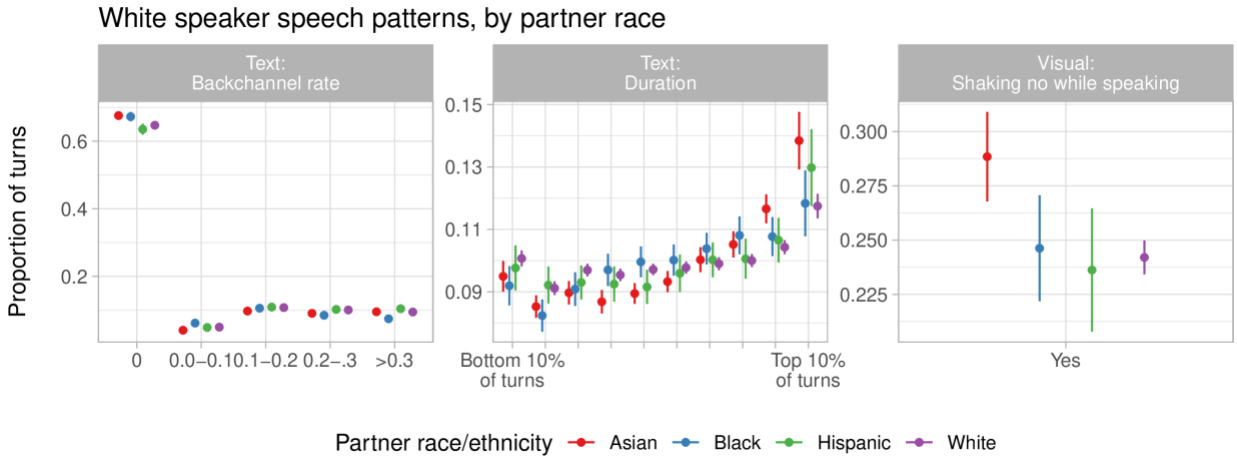

Fig. S.14. Behavior of White participants when assigned to Asian, Black, Hispanic, and White partners. Each panel depicts the engagement patterns of White participants assigned to Asian (red), Black (blue), Hispanic (green), or White (purple) partners on a turn-level characteristic. Horizontal axes denote categories of turn-level characteristics, defined in terms of feature deciles. The vertical position of each point indicates the average proportion of turns in a category. Distributions are presented only for features in which the null hypothesis of distributional equality is rejected at the 0.05 level after multiple-testing adjustment.

Table S.6. Statistical significance of differences in behavior of a participant group, contrasting members assigned to out-group partners versus in-group partners. Each row reports the difference in the behavior of a group of participants toward out-group partners, as compared to in-group partners. The first column specifies the participant group for which conversational behavior is being analyzed. In gender analyses, abbreviations indicate Female and Male groups; in age analyses, Youngest, Middle, and Oldest tertile; in race/ethnicity analyses, Asian, Black, Hispanic, and White. The second column specifies the comparison of partner groups in an abbreviated "X - Y"; in each comparison, the first letter (here, "X") represents the out-group abbreviation and the second ("Y") always corresponds to the participant group being analyzed. Subsequent columns report differences in average conversational behavior toward out-group partners (compared to in-group partners), 95% confidence intervals, and multiple-testing adjusted *p* values for the difference in expectation. To conserve space, only differences in expectation significant at the 95% level after multiple-testing adjustment are reported; note that multiple-testing adjustment accounts for 19 features and a total of 17 partner-group contrasts, totaling 323 analyses; these include additional education and ideology analyses for which no significant difference was found. Winsorized differences are reported for unbounded features.

Participant Group	Partner Contrast	Feature	Diff.	95% CI	p_{adj} (mean)
		Gender-based analyses			
F	M - F	Interval after prior turn	0.0263	[0.0111, 0.0414]	0.018
F	M - F	Pitch	-3.402	[-5.6571, -1.1470]	0.047
F	M - F	Pitch S.D.	-1.213	[-1.8446, -0.5815]	0.007
M	F - M	Facial happiness (speaking)	0.033	[0.0120, 0.0540]	0.038
M	F - M	Facial happiness (listening)	0.034	[0.0127, 0.0554]	0.038
		Age-based analyses			
Y	M - Y	Duration	-0.5208	[-0.8331, -0.2085]	0.027
Y	O - Y	Interval after prior turn	-0.0411	[-0.0660, -0.0161]	0.029
Y	O - Y	Duration	-0.7693	[-1.0665, -0.4722]	<0.001
Y	O - Y	Backchannel rate	0.0072	[0.0026, 0.0119]	0.038
Y	O - Y	Shaking no (speaking)	-0.0338	[-0.0553, -0.0123]	0.038
Y	O - Y	Nodding yes (listening)	0.0365	[0.0122, 0.0608]	0.047
M	O - M	Interval after prior turn	-0.0375	[-0.0622, -0.0127]	0.047
O	M - O	Duration	0.7401	[0.3957, 1.0845]	0.002
O	Y - O	Interval after prior turn	0.045	[0.0234, 0.0666]	0.002
O	Y - O	Duration	1.4652	[1.1082, 1.8222]	<0.001
O	Y - O	Shaking no (speaking)	0.0381	[0.0131, 0.0630]	0.047
O	Y - O	Pitch S.D.	2.4611	[1.2109, 3.7113]	0.005
		Race/ethnicity-based analyses			
W	A - W	Duration	0.7511	[0.4451, 1.0572]	<0.001
W	A - W	Cosine similarity to prior	0.0059	[0.0025, 0.0092]	0.018
W	A - W	Euclidean dist. to prior	-0.0054	[-0.0085, -0.0023]	0.018
W	A - W	Shaking no (speaking)	0.0434	[0.0200, 0.0668]	0.010
W	B - W	Backchannel rate	-0.0128	[-0.0170, -0.0085]	<0.001

REFERENCES AND NOTES

1. H. H. Clark, *Arenas of Language Use* (University of Chicago Press, 1992).

2. N. J. Enfield, *How We Talk: The Inner Workings of Conversation* (Basic Books, 2017).

3. M. J. Pickering, S. Garrod, *Understanding Dialogue: Language Use and Social Interaction* (Cambridge Univ. Press, 2021).

4. H. Sacks, E. A. Schegloff, G. Jefferson, A simplest systematics for the organization of turn-taking for conversation, in *Studies in the Organization of Conversational Interaction*, J. Schenkein, Ed. (Academic Press, 1978), pp. 7–55.

5. M. Tomasello, *Constructing a Language: A Usage-based Theory of Language Acquisition* (Harvard Univ. Press, 2003).

6. M. C. Bateson, Mother-infant exchanges: The epigenesis of conversational interaction. *Ann. N. Y. Acad. Sci.* **263**, 101–113 (1975).

7. C. Trevarthen, K. J. Aitken, Infant intersubjectivity: Research, theory, and clinical applications. *J. Child Psychol. Psychiatry* **42**, 3–48 (2001).

8. S. C. Levinson, J. Holler, The origin of human multi-modal communication. *Philos. Trans. R Soc. B Biol. Sci.* **369**, 20130302 (2014).

9. *S.* Pika, R. Wilkinson, K. H. Kendrick, S. C. Vernes, Taking turns: Bridging the gap between human and animal communication. *Proc. R. Soc. B* **285**, 20180598 (2018).

10. J. Henrich, *The Secret of Our Success* (Princeton Univ. Press, 2015).

11. E. Herrmann, J. Call, M. V. Hernández-Lloreda, B. Hare, M. Tomasello, Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science* **317**, 1360–1366 (2007).

12. R. Dunbar, *Grooming, Gossip, and the Evolution of Language* (Harvard Univ. Press, 1998).

13. R. I. Dunbar, Gossip in evolutionary perspective. *Rev. Gen. Psychol.* **8**, 100–110 (2004).

14. J. Holt-Lunstad, T. F. Robles, D. A. Sbarra, Advancing social connection as a public health priority in the United States. *Am. Psychol.* **72**, 517–530 (2017).

15. A. Milek, E. A. Butler, A. M. Tackman, D. M. Kaplan, C. L. Raison, D. A. Sbarra, S. Vazire, M. R. Mehl, "Eavesdropping on happiness" revisited: A pooled, multisample replication of the association between life satisfaction and observed daily conversation quantity and quality. *Psychol. Sci.* **29**, 1451–1462 (2018).

16. E. Dinan, V. Logacheva, V. Malykh, A. Miller, K. Shuster, J. Urbanek, D. Kiela, A. Szlam, I. Serban, R. Lowe, S. Prabhumoye, A. W. Black, A. Rudnicky, J. Williams, J. Pineau, M. Burtsev, J. Weston, The second conversational intelligence challenge (convai2), in *The NeurIPS'18 Competition* (Springer Cham, 2020), pp. 187–208.

17. A. Ram, R. Prasad, C. Khatri, A. Venkatesh, R. Gabriel, Q. Liu, J. Nunn, B. Hedayatnia, M. Cheng, A. Nagar, E. King, K. Bland, A. Wartick, Y. Pan, H. Song, S. Jayadevan, G. Hwang, A. Pettigrue, Conversational AI: The science behind the alexa prize. arXiv:1801.03604 [cs.AI] (11 January 2018).

18. S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, K. Shuster, E. M. Smith, Y. Boureau, J. Weston, Recipes for building an open-domain chatbot. arXiv:2004.13637 [cs.CL] (28 April 2020).

19. A. Anderson, M. Bader, E. Gurman Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. Mcallister, J. Miller, C. Sotillo, H. S. Thompson, R. Weinert, The HCRC map task corpus. *Lang. Speech* **34**, 351–366 (1991).

20. J. J. Godfrey, E. C. Holliman, J. McDaniel, SWITCHBOARD: Telephone speech corpus for research and development, in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing* (IEEE, 1992), pp. 517–520.

21. S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv:1810.02508 [cs.CL] (5 October 2018).

22. S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, L.-W. Ku, Emotionlines: An emotion corpus of multi-party conversations. arXiv:1802.08379 [cs.CL] (23 February 2018).

23. N. Barros, E. Churamani, H. Lakomkin, H. Sequeira, A. Sutherland, S. Wermter, The OMG-emotion behavior dataset, in *Proceedings of the International Joint Conference on Neural Networks* (IEEE, 2018), pp. 1408–1414.

24. I. V. Serban, R. Lowe, P. Henderson, L. Charlin, J. Pineau, A survey of available corpora for building data-driven dialogue systems. arXiv:1512.05742 [cs.CL] (17 December 2015).

25. S. R. Bowman, G. Angeli, C. Potts, C. D. Manning, A large annotated corpus for learning natural language inference. arXiv:1508.05326 [cs.CL] (21 August 2015).

26. J. Devlin, M. W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [cs.CL] (11 October 2018).

27. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 770–778.

28. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space. arXiv:1301.3781 [cs.CL] (16 January 2013).

29. R. Munro, S. Bethard, V. Kuperman, V. Tzuyin Lai, R. Melnick, C. Potts, T. Schnoebelen, H. Tily, Crowdsourcing and language studies: The new generation of linguistic data, in *NAACL Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk* (Association for Computational Linguistics, 2010), pp. 122–130.

30. The ManyBabies Consortium, Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Adv. Methods Pract. Psychol. Sci.* **3**, 24–52 (2020).

31. Many Primates, D. M. Altschul, M. J. Beran, M. Bohn, J. Call, S. DeTroy, S. J. Duguid, C. L. Egelkamp, C. Fichtel, J. Fischer, M. Flessert, D. Hanus, D. B. M. Haun, L. M. Haux, R. A. Hernandez-Aguilar, E. Herrmann, L. M. Hopper, M. Joly, F. Kano, S. Keupp, A. P. Melis, A. M. Rodrigo, S. R. Ross, A. Sánchez-Amaro, Y. Sato, V. Schmitt, M. K. Schweinfurth, A. M. Seed, D. Taylor, C. J. Völter, E. Warren, J. Watzek, Establishing an infrastructure for collaboration in primate cognition research. *PLOS ONE* **14**, e0223675 (2019).

32. N. A. Coles, J. K. Hamlin, L. L. Sullivan, T. H. Parker, D. Altschul, Build up big-team science. *Nature* **601**, 505–507 (2022).

33. T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. de Ruiter, K.-E. Yoon, S. C. Levinson, Universals and cultural variation in turn-taking in conversation. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10587–10592 (2009).

34. S. C. Levinson, Turn-taking in human communication—Origins and implications for language processing. *Trends Cogn. Sci.* **20**, 6–14 (2016).

35. M. Heldner, J. Edlund, Pauses, gaps and overlaps in conversations. *J. Phon.* **38**, 555–568 (2010).

36. J. P. de Ruiter, H. Mitterer, N. J. Enfield, Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language* **82**, 515–535 (2006).

37. C. Riest, A. B. Jorschick, J. P. de Ruiter, Anticipation in turn-taking: Mechanisms and information sources. *Front. Psychol.* **6**, 89 (2015).

38. S. Bögels, F. Torreira, Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *J. Phon.* **52**, 46–57 (2015).

39. L. Magyari, Predictions in conversation, in *A Life in Cognition*: *Studies in Cognitive Science in Honor of Csaba Pléh*, J. Gervain, G. Csibra, K. Kovács, Eds. (Springer Cham, 2022), pp. 59–75.

40. S. C. Levinson, F. Torreira, Timing in turn-taking and its implications for processing models of language. *Front. Psychol.* **6**, 731 (2015).

41. L. Ten Bosch, N. Oostdijk, L. Boves, On temporal aspects of turn taking in conversational dialogues. *Speech Commun.* **47**, 80–86 (2005).

42. R. E. Corps, B. Knudsen, A. S. Meyer, Overrated gaps: Inter-speaker gaps provide limited information about the timing of turns in conversation. *Cognition* **223**, 105037 (2022).

43. N. G. MacLaren, F. J. Yammarino, S. D. Dionne, H. Sayama, M. D. Mumford, S. Connelly, R. W. Martin, T. J. Mulhearnc, E. M. Todd, A. Kulkarnid, Y. Caoa, G. A. Ruark, Testing the babble hypothesis: Speaking time predicts leader emergence in small groups. *Leadersh. Q.* **31**, 101409 (2020).

44. M. S. Mast, Dominance as expressed and inferred through speaking time: A meta-analysis. *Hum. Commun. Res.* **28**, 420–450 (2002).

45. A. Hepburn, G. B. Bolden, *Transcribing for Social Research* (Sage, 2017).

46. J. P. Simmons, L. D. Nelson, U. Simonsohn, False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).

47. J. B. Bavelas, J. Gerwing, The listener as addressee in face-to-face dialogue. *Int. J. List.* **25**, 178–198 (2011).

48. R. Gardner, *When Listeners Talk* (John Benjamins Publishing Company, 2001).

49. J. B. Bavelas, L. Coates, T. Johnson, Listeners as co-narrators. *J. Pers. Soc. Psychol.* **79**, 941–952 (2000).

50. J. Tolins, J. E. F. Tree, Addressee backchannels steer narrative development. *J. Pragmat.* **70**, 152–164 (2014).

51. G. Jefferson, Caveat speaker: Preliminary notes on recipient topic-shift implicature. *Res. Lang. Soc. Interact.* **26**, 1–30 (1993).

52. G. Brown, G. Yule, *Discourse Analysis* (Cambridge Univ. Press, 1983).

53. H. P. Grice, Logic and conversation, in *Speech Acts* (Brill, 1975), pp. 41–58.

54. P. Brown, S. C. Levinson, *Politeness: Some Universals in Language Usage* (Cambridge Univ. Press, 1987), vol. 4.

55. S. Albert, J. P. de Ruiter, Repair: The interface between interaction and cognition. *Top. Cogn. Sci.* **10**, 279–313 (2018).

56. N. D. Duran, A. Paxton, R. Fusaroli, ALIGN: Analyzing linguistic interactions with generalizable techNiques—A Python library. *Psychol. Methods* **24**, 419–438 (2019).

57. C. Dideriksen, M. H. Christiansen, K. Tylén, M. Dingemanse, R. Fusaroli, Quantifying the interplay of conversational devices in building mutual understanding. PsyArXiv (12 October 2020). https://psyarxiv.com/a5r74/.

58. D. S. Berry, J. S. Hansen, Positive affect, negative affect, and social interaction. *J. Pers. Soc. Psychol.* **71**, 796–809 (1996).

59. L. A. Clark, D. Watson, Mood and the mundane: Relations between daily life events and self-reported mood. *J. Pers. Soc. Psychol.* **54**, 296–308 (1988).

60. L. C. Hawkley, J. T. Cacioppo, Loneliness matters: A theoretical and empirical review of consequences and mechanisms. *Ann. Behav. Med.* **40**, 218–227 (2010).

61. N. Epley, M. Kardas, X. Zhao, S. Atir, J. Schroeder, Undersociality: Miscalibrated social cognition can inhibit social connection. *Trends Cogn. Sci.* **26**, 406–418 (2022).

62. N. Epley, J. Schroeder, Mistakenly seeking solitude. *J. Exp. Psychol. Gen.* **143**, 1980–1999 (2014).

63. J. Schroeder, D. Lyons, N. Epley, Hello, stranger? Pleasant conversations are preceded by concerns about starting one. *J. Exp. Psychol. Gen.* **151**, 1141–1153 (2021).

64. E. J. Boothby, G. Cooney, G. M. Sandstrom, M. S. Clark, The liking gap in conversations: Do people like us more than we think? *Psychol. Sci.* **29**, 1742–1756 (2018).

65. G. Cooney, E. J. Boothby, M. Lee, The thought gap after conversation: Underestimating the frequency of others' thoughts about us. *J. Exp. Psychol. Gen.* **151**, 1069–1088 (2022).

66. A. Mastroianni, G. Cooney, E. J. Boothby, A. G. Reece, The liking gap in groups and teams. *Organ. Behav. Hum. Decis. Process.* **162**, 109–122 (2021).

67. E. M. Templeton, L. J. Chang, E. A. Reynolds, M. D. C. LeBeaumont, T. Wheatley, Fast response times signal social connection in conversation. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2116915119 (2022).

68. B. L. Smith, B. L. Brown, W. J. Strong, A. C. Rencher, Effects of speech rate on personality perception. *Lang. Speech* **18**, 145–152 (1975).

69. N. Miller, G. Maruyama, R. J. Beaber, K. Valone, Speed of speech and persuasion. *J. Pers. Soc. Psychol.* **34**, 615–624 (1976).

70. N. A. Murphy, J. A. Hall, C. R. Colvin, Accurate intelligence assessments in social interactions: Mediators and gender effects. *J. Pers.* **71**, 465–493 (2003).

71. M. Kusner, Y. Sun, N. Kolkin, K. Weinberger, From word embeddings to document distances, in *International Conference on Machine Learning* (PMLR, 2015), pp. 957–966.

72. K. Song, X. Tan, T. Qin, J. Lu, T. Y. Liu, MPNet: Masked and permuted pre-training for language understanding. *Adv. Neural Inf. Process. Syst.* **33**, 16857–16867 (2020).

73. N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv:1908.10084 [cs.CL] (27 August 2019).

74. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692 [cs.CL] (2019).

75. M. Florentine, Loudness, in *Loudness*, M. Florentine, A. N. Popper, R. R. Fay, Eds. (Springer, 2011), pp. 1–15.

76. L. E. Marks, M. Florentine, Measurement of loudness, part I: Methods, problems, and pitfalls, in *Loudness*, M. Florentine, A. N. Popper, R. R. Fay, Eds. (Springer, 2011), pp. 17–56.

77. R. A. Page, J. L. Balloun, The effect of voice volume on the perception of personality. *J. Soc. Psychol.* **105**, 65–72 (1978).

78. A. C. Weidman, J. L. Tracy, Picking up good vibrations: Uncovering the content of distinct positive emotion subjective experience. *Emotion* **20**, 1311–1331 (2020).

79. E. Diener, R. J. Larsen, S. Levine, R. A. Emmons, Intensity and frequency: Dimensions underlying positive and negative affect. *J. Pers. Soc. Psychol.* **48**, 1253–1265 (1985).

80. R. Reisenzein, Pleasure-arousal theory and the intensity of emotions. *J. Pers. Soc. Psychol.* **67**, 525–539 (1994).

81. J. Sonnemans, N. H. Frijda, The structure of subjective emotional intensity. *Cogn. Emot.* **8**, 329–350 (1994).

82. S. R. Livingstone, F. A. Russo, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE* **13**, e0196391 (2018).

83. D. E. King, Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.* **10**, 1755–1758 (2009).

84. A. Mollahosseini, B. Hasani, M. H. Mahoor, AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **10**, 18–31 (2017).

85. L. F. Barrett, Solving the emotion paradox: Categorization and the experience of emotion. *Pers. Soc. Psychol. Rev.* **10**, 20–46 (2006).

86. U. Hess, C. Blaison, K. Kafetsios, Judging facial emotion expressions in context: The influence of culture and self-construal orientation. *J. Nonverbal Behav.* **40**, 55–64 (2016).

87. T. Stivers, J. Sidnell, *The Handbook of Conversation Analysis* (John Wiley & Sons, 2012).

88. E. Stokoe, *Talk: The Science of Conversation* (Hachette UK, 2018).

89. D. Jurafsky, D. Tolinsky, R. Ranganath, D. McFarland, Extracting social meaning: Identifying interactional style in spoken conversation, in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics, 2009), pp. 638–664

90. J. F. Dovidio, K. Kawakami, S. L. Gaertner, Implicit and explicit prejudice and interracial interaction. *J. Pers. Soc. Psychol.* **82**, 62–68 (2002).

91. M. Sen, O. Wasow, Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annu. Rev. Polit. Sci.* **19**, 499–522 (2016).

92. J. Henrich, S. J. Heine, A. Norenzayan, Most people are not WEIRD. *Nature* **466**, 29–29 (2010).

93. G. Cooney, A. Mastroianni, N. Abi-Esber, A. W. Brooks, The many minds problem: Disclosure in dyadic versus group conversation. *Curr. Opin. Psychol.* **31**, 22–27 (2020).

94. R. L. Moreland, Are dyads really groups? *Small Group Res.* **41**, 251–267 (2010).

95. T. Stivers, Is conversation built for two? The partitioning of social interaction. *Res. Lang. Soc. Interact.* **54**, 1–19 (2021).

96. E. Jolly, L. J. Chang, Gossip drives vicarious learning and facilitates social connection. *Curr. Biol.* **31**, 2539–2549.e6 (2021).

97. K. Greene, V. J. Derlega, A. Mathews, Self-disclosure in personal relations, in *The Cambridge Handbook of Personal Relations*, A. L. Vangelisti, D. Perlman, Eds. (Cambridge Univ. Press, 2006), pp. 409–427.

98. D. Wilkes-Gibbs, H. H. Clark, Coordinating beliefs in conversation. *J. Mem. Lang.* **31**, 183–194 (1992).

99. D. Knox, C. Lucas, A dynamic model of speech for the social sciences. *Am. Polit. Sci. Rev.* **115**, 649–666 (2021).

100. C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, C. Potts, A computational approach to politeness with application to social factors. arXiv:1306.6078 [cs.CL] (25 June 2013).

101. M. Yeomans, A. Kantor, D. Tingley. The politeness package: Detecting politeness in natural language. *R Journal* **10** 489–502 (2018).

102. Y. Jadoul, B. Thompson, B. De Boer, Introducing parselmouth: A python interface to praat. *J. Phon.* **71**, 1–15 (2018).

103. B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: Audio and music signal analysis in python, in *Proceedings of the 14th Python in Science Conference* (2015), vol. 8, pp. 18–25.

104. P. Mermelstein, Distance measures for speech recognition, psychological and instrumental. *Pattern Recognit. Artif. Intell.* **116**, 374–388 (1976).

105. M. Farrús, J. Hernando, P Ejarque, Jitter and shimmer measurements for speaker recognition, in *Proceedings of the 8th Annual Conference of the International Speech Communication Association* (International Speech Communication Association, 2007), pp. 778–781

106. P. Boersma, D. Weenink, Praat: Doing phonetics by computer [Computer program], version 6.2.08 (2022); www.praat.org/.

107. SPTK Working Group, Speech Signal Processing Toolkit (SPTK) (2017); http://sp-tk.sourceforge.net.

108. N. Dehak, P. Dumouchel, P. Kenny, Modeling prosodic features with joint factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process.* **15**, 2095–2103 (2007).

109. J. Howard, R. Thomas, S. Gugger, fastai. GitHub (2018).

110. S. Dutta, S. Datta, A rank-sum test for clustered data when the number of subjects in a group within a cluster is informative. *Biometrics* **72**, 432–440 (2016).

111. C. B. Hilton, C. J. Moser, M. Bertolo, H. Lee-Rubin, D. Amir, C. M. Bainbridge, J. Simson, D. Knox, L. Glowacki, E. Alemu, A. Galbarczyk, G. Jasienska, C. T. Ross, M. Neff, A. Martin, L. K. Cirelli, S. E. Trehub, J. Song, M. Kim, A. Schachner, T. A. Vardy, Q. D. Atkinson, A. Salenius, J. Andelin, J. Antfolk, P. Madhivanan, A. Siddaiah, C. D. Placek, G. Deniz Salali, S. Keestra, M. Singh, S. A. Collins, J. Q. Patton, C. Scaff, J. Stieglitz, S. Ccari Cutipa, C. Moya, R. R. Sagar, M. Anyawire, A. Mabulla, B. M. Wood, M. M. Krasnow, S. A. Mehr, Acoustic regularities in infant-directed speech and song across cultures. *Nat. Hum. Behav.* **6**, 1545–1556 (2022).

112. C. Yale, A. B. Forsythe, Winsorized regression. *Technometrics* **18**, 291–300 (1976).

113. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B* **57**, 289–300 (1995).