Perceptions of Interaction Styles: An Exploration using Speaker-Typical Utterances

Nigel G. Ward[1] and Georgina Bugarini[1] and Harm Lameris[1,2]
[1] University of Texas at El Paso, El Paso, Texas, USA
[2] KTH Royal Institute of Technology, Stockholm, Sweden
nigelward@acm.org, georgina.bugarini20@yahoo.com, lameris@kth.se

Abstract

Different people have different dialog interaction styles, but our understanding of how these are perceived is limited. We approach this problem by making two working assumptions: that interaction style perceptions are largely just the result of utterance-by-utterance perceptions of a speaker's pragmatic intent, and that pragmatically typical utterances for a speaker can be informative regarding their overall interaction style. We accordingly applied a recently-developed vector-space model of pragmatic functions to identify speaker-typical utterances. We found, among other things, indications that people are able to infer useful style information from single utterances, and that the space of interaction styles is very broad.

1 Motivation

In terms of dialog interaction style, we may perceive people variously as domineering, laid-back, no-nonsense, chatty, good listeners, and so on. While the notion of "interaction style" has no standard definition, we here, inspired by (Tannen, 1980), consider an interaction style to be constellation of linguistic features and interactive behavior propensities that enable others to predict a person's likely behaviors and responses in conversational interaction, and thereby infer how to effectively interact with them.

Studies of conversation and dialog currently tend to seek general findings by abstracting away from individual differences. While this has been a productive strategy, the field may be approaching readiness to start exploring the effects of such differences, specifically in interaction styles. Doing so could benefit both empirical studies of conversation phenomena and the endeavor of building dialog systems better able to satisfy individual user preferences.

However, there have been very few explorations of interaction style differences and how they are perceived. Here we approach this question in a novel way, through speaker-typical utterances, leveraging a recently-developed vector-space model of pragmatic functions, and building on research on perceptions of thin slices of behavior.

2 Use Cases

One of our motivations for exploring interaction styles is a practical one: giving someone advance knowledge of a new interlocutor's interaction style may be useful.

In daily life, talking with other people about the interaction styles of third parties is not uncommon, and this can sometimes enable us to enter a conversation with a new interlocutor with useful expectations. Such information sharing often occurs informally, when we happen to meet someone who knows the other person. However, in modern society, the number of potential interaction partners is huge, and we generally lack connections who can provide such information.

We accordingly conceived of an alternative way to convey such information: through a single typical utterance. We suspected that a truly typical utterance could serve as a "distillation" of a substantial sample of behavior, helpful to enable a listener to quickly grasp enough about the person to feel better prepared to enter an interaction, or to predict the likely value of choosing to interact with that person at all. Figure 1 illustrates the concept. More specific use cases include:

1. Preparation for a Conversation. Before meeting someone for the first time, we commonly look them up on the web or social media, to get a feel for what they will be like to talk to, but this only helps a little. If instead their interaction style could be represented by a single typical utterance, we could quickly listen to that first, to get a rough idea of who we will be dealing with and what conversa-

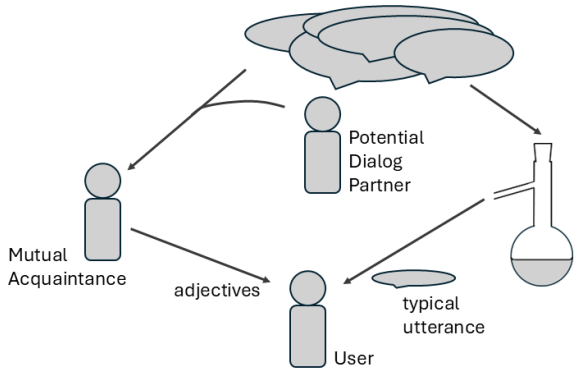

Figure 1: Our Applications Concept: A user, bottom, interested in potentially interacting with someone new, top, might learn something about them from a mutual acquaintance, left. As an alternative, we propose to support the user by providing a typical utterance that is automatically extracted from a corpus of speech by the potential dialog partner, right.

tional style might work best with them: directive, relaxed, businesslike, friendly, etc.

2. Partner Selection. Informed choice of teammates, tutors, romantic partners, consultants, and other interaction partners is important, for individuals and for society. There are many established processes and resource types to support this (Lykourentzou et al., 2017; Palea et al., 2024), but these may give little idea of what the person would be actually like to interact with (Azuma et al., 2025). If a person's authentic interaction style could be represented by a single utterance, then people could potentially match up faster with more suitable interaction partners. As a special case, users may want similarly to select among variant personas for a dialog system, or use an exemplar utterance to indicate their ideal for what the system should be like (Lilley et al., 2025; Ostrowski et al., 2022; Metcalf et al., 2019).

3. Helping Patients and Clients. Insofar as people with autism, and possibly other conditions, have characteristic interaction styles (Scheeren et al., 2012; Matyjek et al., 2026), typical utterances could be useful in clinical work. We recently surveyed 18 speech-language pathologists regarding some possible ways in which technology could support their workflow, inspired by (Klatte et al., 2022) and (Ward et al., 2025). We asked them to rate the idea that "clinicians might appreciate the ability to quickly find and hear a 'typical' utterance by a speaker of interest, as that could serve as a useful summary of the properties of many utterances. This could be useful for providing examples in assessment reports." Of the 18 participants, 11 agreed or strongly agreed that this would be a use case worth pursuing. Their reasons included, beyond the idea of identifying examples to include in reports, helping parents and children understand the diagnosis and recommendations, as in: "The 'typical' utterance may open up conversation regarding what the expectations are for both listener and speaker as well as clarify purposes of communicative behaviors and why behaviors are expected or considered 'atypical'."

3 Research Questions

We accordingly set out to explore the space of interaction styles and how they are perceived. As an exploration, we did not have specific expectations for what we might find, but our research questions included, vaguely:

RQ 1 What interaction styles exist?

RQ 2 Do observers agree on what styles are present in speech samples? and

RQ 3 Can observers perceive interaction style information from even just a single typical utterance?

We also aimed to lay the groundwork for building out the use cases. The associated research questions were:

RQ 4 how best to select typical utterances, and

RQ 5 whether people find them of value for rapid identification of interaction styles.

4 Related Research

We survey related research around five themes: interaction styles, thin slices, inferring user properties from speech data, pragmatic functions in conversation, and modeling typicality.

Work on interaction styles, starting from Tannen's (Tannen, 1980, 1990) seminal work on conversation styles, has identified many dimensions of variation. The general focus of this work has been identifying cultural, group, or demographic differences, rather than characterizing individuals' styles (Geertzen, 2015), as surveyed by (Ward and Avlia, 2023). An important observation is that interaction styles are not stable properties of people, but behavior instead varies based on factors such as the interlocutor and the context (Selting, 1989; Zhao et al., 2026; Ward and Avlia, 2023). However,

in order to make progress identifying styles in the first place, we here accept the limitation of only addressing those aspects of interaction styles that are relatively stable for each speaker.

There is also some computational work on this topic. Some work treats only a small number of styles, such as awkward, friendly, or flirtatious (Ranganath et al., 2013). Dimensional models, in contrast, can describe greater variety, and naturally reflect the fact that various nonverbal features tend to co-occur (Tannen, 1980; Grothendieck et al., 2011; Ward and Avlia, 2023). In such models, a person's style is described in terms of its values on the dimensions, or, equivalently as a vector in a space spanned by those dimensions. Proposed dimensions cover attributes such as a content focus versus an interlocutor orientation, being engaged, taking a fatalistic/accepting stance, being upbeat, and favoring the listener role.

Work on "thin slices" in the fields of social and personality psychology has shown that observations of short behavior samples often enable listeners to make social-construct judgments that are highly predictive of judgments made on the basis of much more data (Murphy and Hall, 2021; Ambady and Rosenthal, 1992). This phenomenon has been seen in predicting, for example: suitability for hiring from interviews, the likelihood of couples staying together, ratings of teacher effectiveness, and speed-dating outcomes. Thin-slicing research has mostly used video data, but judgments from audio alone are sometimes even more informative (Kraus, 2017). Thin-slice observation is effective for slices of various sizes (Ambady and Rosenthal, 1992; Knops and Hagen, 1989; McAleer et al., 2014; Hall et al., 2014; Holliday, 2023; Lavan et al., 2024; Chakravarthula et al., 2021; Wang et al., 2021), but previous work has not explored the utility of single-utterance slices. To date, the question of whether some slices are more informative than others (Carney et al., 2007; Ingold et al., 2024; Le et al., 2024; Azuma et al., 2025; Wang et al., 2021; Hall et al., 2021), has been only sporadically examined, and only as a question of pure scientific interest. Here, motivated by the use cases, we are working with uncommonly short slices, so effective slice selection is an important issue.

In speech technology, there is a long tradition of work on inferring speaker properties from speech, such as personality traits, communication skills, and interaction styles (Schultz, 2007; Li et al., 2024; Nguyen and Gatica-Perez, 2015; Völkel et al., 2020; Gallardo and Weiss, 2018; van Rijn et al., 2024; Ranganath et al., 2013; Tannen, 1980; Grothendieck et al., 2011; Ward and Avlia, 2023; Okada et al., 2016). This work largely relies on the same types of features that underpin the success of thin slicing: fast, subtle nonverbal signals, often prosodic, that are generally below the level of conscious awareness (Pentland, 2007; Vinciarelli et al., 2011). Such classifiers often perform very well, at least on curated data.

A related thread of research has explored perceptions of voices and interaction styles for automated agents. Some work, inspired by the problem of selecting personas or voices for spoken dialog systems, has explored the relationships between perceived voice properties and higher-level speaker characteristics (Gallardo and Weiss, 2018; van Rijn et al., 2024), including interpersonal characteristics such as "likable, attractive, competent, childish," and so on. However we doubt that any small finite set of adjectives will be adequate for describing the rich variety of human interaction styles, as suggested by the work of (Völkel et al., 2020). Further, while personality attributes are often exploited in agent design (Kim et al., 2019), (Völkel et al., 2020) shows that the well-studied Big 5 personality attributes are not adequate for describing the interaction style elements attributes that are most important to conversational interaction. They present a more complete personality model with dozens of associated descriptive adjectives (mostly organized into 10 factors), but note that, in the end, even this will not be enough for characterizing interaction partners in general.

Our understanding of the pragmatic functions employed in interaction has grown with detailed analysis of conversation behaviors (Clark, 1996; Couper-Kuhlen and Selting, 2018): there are many things that people do with language beyond just conveying information. Beyond the well-known speech acts and dialog acts — question, answer, command, request, confirm, acknowledge, and so on — people also do things like mark something as new information, yield the turn, invite a backchannel, assess something positively, cue action, ground an ambiguous referring expression, signal a topic change, indicate confidence, defer to the interlocutor's superior knowledge, show empathy, and so on. Many, if not most, utterances in daily conversation simultaneously convey several such functions, in goal- and situation-tailored combinations. There are clear connections between pragmatics-related

behaviors and social meaning, but research at the intersection has been scant (Beltrama, 2020).

Regarding the notion of typicality, an influential perspective defines the most typical member of a set to be the one that is most similar, on average, to all others in that set (Belohlavek and Mikula, 2024). This idea works well with embeddings: mathematically, if we use a vector representation that behaves as a metric space, then the utterance that is nearest to the centroid (the average), will be closest, on average, to all the other utterances in the set (Rossiello et al., 2017). Many vector spaces have this property, including, in practice, most embeddings. Averages have seen some use in speech processing (Oplustil Gallegos et al., 2020; Fontaine et al., 2017; Bruckert et al., 2010), and caricatures have been used in voice manipulations (López et al., 2013), but neither has been considered for utterances.

Thus, despite a good body of related work, our understanding of interaction styles is still limited.

5 Methods and Quantitative Results

Our research strategy was to initially focus on the use cases. Among other advantages, this gave our participants a sense of purpose and focus: rather than try to involve them in blank-slate contemplations, we engaged them in our aim of creating utility. They were particularly enthusiastic about the potential to help children through Use Case 3.

5.1 Model

We needed a model able to find utterances that are typical in a specific way: in terms of interaction styles. As already noted, we chose to try, as a proxy, an embedding that represents pragmatic-function-related behavior. As a side note, we might have tried many feature sets, relating to turn taking, lexical content, or prosody (Grothendieck et al., 2011; Miehle et al., 2020, 2022; Ward and Avlia, 2023; Wong and Ginzburg, 2018), but we opted for an embedding (Wang et al., 2024), expecting it to encode much more information.

Given this choice, only one model was available: our own previous work (Ward et al., 2024). In technical terms, this model is simply a subspace of the HuBert Large embedding (Hsu et al., 2021), consisting of 103 features (dimensions), selected for utility for predicting human judgments of pragmatic similarity among utterance pairs. Importantly for our purposes, this may capture more detail than

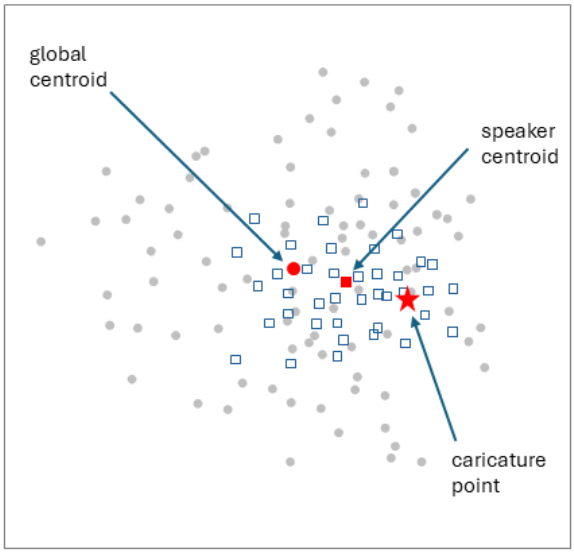

Figure 2: The Caricature Concept. Squares represent utterances from the speaker being modeled, and dots represent utterances from other speakers.

classic dialog-act taxonomies, reflecting not only such top-level pragmatic distinctions as question versus statement, but many more nuances of what people are trying to do in conversation and how they do it.

We must digress to acknowledge here that not all elements of style are utterance-internal, with turn-taking dynamics only the most obvious additional factor to consider (Roberts and Francis, 2013; Blohm and Barthel, 2025). We nevertheless limit our scope here to what we can learn from utterances without context. This is not unreasonable, because we are working with speech data from conversations, in which even a single utterances can portray many aspects of interaction style, especially since 5 second clips usually contain enough information to infer something about what came just before. For example, from the following utterance (Switchboard 3003@0:05.7)

> *well, no, it's probably going to be a very old used one; my husband had an accident . . .*

it's easy to infer (in part from the prosody) that the topic is buying cars, that the speaker had earlier said that she was in the market for one, and that the interlocutor had made a comment or asked a question that presupposed that she would be shopping for a new one and looking forward to it.

For choosing typical utterances we developed a way to find "caricatures." Given the vector space, a

straightforward implementation of typicality would be to compute the average of all of the speaker's utterances, namely the centroid, and return the utterance closest to it. (Here, based on a pilot study, we chose to compute closeness using Euclidean distances rather than cosines.) However, most people are rather alike: the average behavior for one person is generally not far from the average population behavior. To convey more information, we decided to try using utterances that exaggerate the differences: caricature utterances. Figure 2 illustrates the concept. Specifically, for finding the caricature, we used C = S + α(S–G), where G is the group centroid and S the speaker centroid. We set α to be 1.5, based on a pilot study.

5.2 Evaluation Methods

The first evaluation instrument, addressing RQ 4, was based on the idea that a typicality model is better if its selections of typical utterances are indeed more speaker-typical according to human judgments. Inspired by previous studies of typicality (Le Mens et al., 2023; Belohlavek and Mikula, 2024), we accordingly had participants "rate the following utterances in terms of most typical to least typical," using a slider with 1 and 5 as the endpoints. We hypothesized that the caricatures would be judged highly typical, and, in particular, more typical than the centroids and the utterances from three other conditions: chosen randomly, chosen as the one closest to the middle of one of the conversations, and chosen as a "first," that is, an utterance that was closest to the beginning of one of the conversations.

The second instrument, for RQ 5, asked the participants to: "On a scale of 1-5, please rate the value of having the most typical utterance of a speaker," where the anchors were "1: no value" and "5: very valuable." There were five sliders, one for the general scenario of meeting a new person (Use Case 1), and four for specific scenarios, of choosing a guide to a new city, a counselor, a teammate, and a conversation partner (Use Case 2). We hypothesized that participants would consider typical utterances to generally have value. We asked these questions after they had heard five candidates for typical utterances and identified the best, so they knew what "the most typical utterance" might be like.

Third, we asked participants to "please provide adjective(s) to describe each utterance in terms of what it reveals about the speaker's conversational style or personality," giving them a list of suggestions but also asking them to add some of their own. The suggestions served to guide them to focus on speaker traits, rather than voice properties or ephemeral intents. Having participants provide descriptors also helped keep them engaged and perceptive.

The procedure had four phases. 1) Participants listened to each of 5 clips and chose some descriptors for each. The five clips were presented in random order. Any participant could request that we replay any clip, any number of times, and participants did this often. 2) They listened to substantial conversational speech data for the speaker and chose more descriptors. 3) They listened again to the 5 clips and made judgments of typicality, again being able to request any number of replays. 4) They made judgments regarding utility. The entire procedure took about 15 minutes and was repeated for each of 11 speakers. The protocol was approved by our Institutional Review Board.

The training of participants was brief. We motivated our interest in typical utterances with the use cases. We told them that we were especially interested in which clips were highest rated, to guide them to consider the task as one of ranking as well as rating. We did not define "typical" nor delimit what might be included in "interaction style," instead allowing them to interpret these terms to mean whatever speaker properties they thought were relevant to the use cases. After the first and second data sets we had group discussions, to help the participants sharpen their awareness and perhaps become more consistent in the judgments they made.

Participants needed to be able to listen closely, make sensitive judgments, reflect on why they made these judgments, and maintain sustained attention. We therefore recruited participants carefully: 10 who all had previously participated in other studies and were comfortable with close listening. All participants were part of our university community, mostly undergraduates, and were compensated generously by local standards: US$70 plus snacks and lunch for coming in for four hours on a Sunday.

5.3 Stimulus Selection and Preparation

Our stimuli were taken from the Switchboard corpus (Godfrey et al., 1992; Deshmukh et al., 1998), because our interest in dialog behaviors demanded a dialog corpus, because we needed a corpus with dialects familiar to our participants, and because

source	average	std. dev.
caricature	3.51	±0.89
first	3.46	±0.87
centroid	3.26	±0.83
random	3.24	±0.97
middle	3.19	±1.06

Table 1: Typicality Ratings for the 5 conditions.

we are interested in the typical behavior of speakers across their interactions with multiple partners.

To find the centroids and caricatures for each speaker, the model used utterances of at least 5 seconds from all conversations involving that speaker, totaling around 25 to 100 minutes of data. It being infeasible to present all of a speaker's conversations to the participants, they only heard 1 minute each from 3 conversations, taken towards the middle of each conversation, each with a different interlocutor.

5.4 Typicality and Value Rating Results

Regarding RQ 4, the caricatures were on average rated more typical than utterances from the other conditions (Table 1). This was significant by a Linear Mixed Model (LMM), where the dependent variable was the log-scaled ratings ($n = 550$). We chose log-scaling because the distribution was skew, with very few low typicality values, and log scaling provided a better model fit than the unmodified ratings. We thus used the model $log(Rating) \sim Condition + (1|JudgeID) + (1|SpeakerID) + (1|OrderID)$ (in Wilkinson notation). We used the condition (the utterance source) as the fixed effect, and the participants ($n = 10$), speakers ($n = 11$), and order of the utterance per speaker ($n = 5$) as the random effects. This yielded a strongly positive coefficient for the caricature condition ($p < .0001$), meaning that the caricature was rated as more typical. While the "first" condition was on average rated slightly less typical than the caricature condition, this difference was not significant ($p = .48$). However, the advantages over the centroid and random utterances were significant, as hypothesized.

Regarding RQ 5, participants felt that typical utterances usually did have value, as hypothesized: of the 550 utility judgments, only 4% were a 1 (no value), and 90% were rated 3 or higher. We also found that the perceived value rating was especially high for the teammate-selection scenario, for which there were zero "no value" judgments, and for the counselor-selection scenario, which had the highest average rating: 4.0 on the 1–5 scale.

5.5 Implications for the Use Cases

We found that caricatures are perceived as speaker-typical, that people can perceive a lot of information from typical utterances, and that they consider this useful. These suggest, in general, potential utility for the use cases. The potential utility is especially high for the scenario of finding someone to serve in a role where interaction skills are paramount, such as a counselor or a teammate. Nevertheless, deployment of any such application would need to come after examination of ethical and practical concerns. Only for some contexts, such as dating sites or institution-internal tracking, is it easy to imagine that people would be willing to have their utterances archived, searched, and displayed in this way. However concerns might be mitigated if the typical utterances were not actual utterances from real conversations, but synthetic ones, perhaps created using style transfer techniques (Wang et al., 2018; Sigurgeirsson and King, 2024), to illustrate the interaction style without violating privacy.

6 Observations and Qualitative Discussion

Returning now to our broader questions, we report what we learned from participants' comments during discussions and from the descriptors they chose.

6.1 The Space of Interaction Styles

First, we note that the concept of interaction style was meaningful to the judges: none reported difficulty with the concept of interaction style, and, with one exception, ascribed at least one descriptor to each clip or segment, and usually two or more.

Regarding RQ 1, we found that participants' concepts of interaction styles were very broad. The most frequently chosen descriptors were among those on the list of suggestions: *friendly, confident, agreeable*, etc., other common descriptors were *thoughtful, calm*, and *engaged*, as summarized in Table 2. However participants also used a long tail of idiosyncratic descriptors, such as *abstract, blustery, cursory, lightweight, modern, puerile, resolute*, and *transparent*, each used only once. The combined number of unique descriptors across the

descriptor	cumulative count for: single utts.	minute-long samples	total
friendly*	49	90	139
confident*	70	50	120
agreeable*	22	78	100
mature*	41	41	82
extroverted*	29	30	59
introverted*	34	25	59
thoughtful	24	21	45
calm	26	16	42
insecure*	28	12	40
engaged	8	32	40
open-minded*	10	28	38
pleasant*	12	24	36
conscientious*	16	20	36
hesitant	17	11	29
closed-minded*	9	19	28
irritating*	23	4	27
uninterested	10	13	23
interested	6	14	20
curious	4	16	20
opinionated	8	12	20
supportive	3	15	18
kind	7	10	17
unsure	8	8	16
attentive	0	14	14
distant	5	9	14
caring	8	6	14
chatty	9	4	13
judgmental	6	7	13
bored	8	5	13
reluctant	9	3	12
happy	3	9	12
humorous	7	5	12
disagreeable*	7	5	12
passionate	5	7	12
dominant	4	7	11
monotone	10	1	11
careful	4	6	10
respectful	2	7	9
polite	2	7	9
careless*	7	2	9
timid	7	2	9
...			
Total	908	1047	1955

Table 2: Descriptors used 9 times or more. * marks those on the list of suggested descriptors.

short utterances and the conversations was 345, after correcting variant spellings. Thus, while a handful of dimensions may explain most of the variance, the idea of boiling everything down to just 10 or 5 or 2 dimensions (Abele and Wojciszke, 2013; McAleer et al., 2014; Su et al., 2025; Yan et al., 2026), would likely be inadequately informative for many purposes (Freeman and Lin, 2025).

6.2 Agreement Among Participants

Regarding RQ 2, judges partly agreed in their perceptions of the speakers. In free discussion this frequently came up as a topic of interest. Consider, for example, clip 4 for speaker 1039 (sw2190@0:43), in which the speaker said

> *well, th- they talked nuclear winter, and, and I suppose that's the ultimate, uh*

The descriptors this garnered, across the 10 judges, were: *jumpy, anxious; close-minded, confident; timid, caring; insecure; mature, agreeable; introverted; thoughtful; uncertain, reluctant; calm*; and *confident, mature*.

Regarding differences, some judges' impressions really differed, as seen by the fact that two used the word *confident*, and two others chose *uncertain* and *timid*. The reasons for such differences were generally hard to discern. Participants never suggested that someone else's judgments might have been mistaken, but there was no tendency to groupthink: during discussion, no one ever expressed a desire to change their descriptors after hearing what someone else thought. Generally, they seemed to accept different perceptions as all valid.

Regarding similarities, many descriptors were quite consistent, for example, here within the set {*introverted, thoughtful, calm, mature*}, and within the set {*jumpy, anxious, timid, insecure, uncertain*}. Overall there seemed to be more similarities than differences, and, in discussion, judges were often pleased to discover that they had agreed in noticing some aspect of the speaker, especially when someone else's descriptor was a particularly apt term for something that they had perceived.

We also investigated how well humans perform the same task, of choosing a typical utterance for a speaker. We wanted to know whether people agreed with each other more than with the model-chosen caricatures. We can answer this questions only approximately with this data, since we do not know which utterances the participants would truly

have felt most typical, had they judged all of those uttered by a speaker. But we could measure agreement over the 5 utterances that they all judged. We found, first, that the average typicality rating of the top picks of other participants was 3.58, slightly higher than the 3.51 for the caricatures, but not significantly so, by an unmatched-pairs t-test. Second, when considering only which utterances were rated highest, we found that the caricature was ranked most typical 24.5% of the time, and the utterance that was selected as most typical by another participant 25.5% of the time. Thus, by both measures, the humans agreed with the caricature-picking algorithm only slightly less than they agreed with each other.

6.3 Reliability of Single-Utterance Perceptions

Regarding RQ 3, Table 2 reveals that some descriptors are much more common after extended listening than after the single utterances, notably *engaged, curious, supportive*, and *attentive*. These all relate to interactive behaviors, which are, of course, rarely evident from utterances without context.

Conversely, some descriptors are more common for the single utterances. Some of these relate to transient speaker states, such as *hesitant*. Others involve negative judgments, notably *monotone* and *irritating*, and including also *disagreeable*, *inflexible*, *distracted*, *sad* and *dominant*. We did a post-hoc comparison of the prevalence of negative descriptors for the utterance samples versus the minute-long samples. There was indeed a bias: they were more common for the single utterances ($p < .00001$ by a post hoc chi-square test). Overall the difference between the negative-descriptor ratios in the two conditions is 10% (29% - 19%), which suggests that judgments based on the single utterances are partly misleading at least 10% of the time, although this ranged from 0% for some participants to 26% for one, probably reflecting variation in interpersonal sensitivity (Shoda, 2011). Learning may have taken place: the excess negativity was 33% for the first speaker judged, but only 7% for the last, suggesting that the bias might fade with more experience, or might be mitigable through training.

7 Future Work

This study represents only a case-study exploration of the research questions, and has many limitations. To support further investigations, we make the judgment data available at https://www.cs.utep.edu/nigel/typical/, along with code for the model and analysis scripts.

Despite the limitations, this work could be a stepping stone to deeper and more focused explorations. One could work to turn our rough list of interaction style descriptors into a proper taxonomy. One could use this to revisit earlier work on group propensities. One could explore the utility of these interaction style descriptors for customizing agent behavior, beyond the usual bland prompting of large language models with "you are a helpful assistant." One could explore how conversation participants' utterance-level behaviors are affected by how they perceive their interlocutor's style.

8 Summary

Our contributions include: 1) Perceptions of interaction styles are diverse. 2) Perceptions are partly shared among observers. 3) Utterances presented without context, as thin slices, have value for forming perceptions of interaction style, 4) While judgments based on such utterances tend to be biased, the bias declines with experience, and 5) There is, indeed, a connection between pragmatic-function usage tendencies and perceived pragmatic styles. We also contribute 6) An algorithm and proof of concept for a new application for speech technology, in which speaker-typical utterances are used to support people needing information about another's interaction style.

Acknowledgments

We thank Julie Kientz and Hyewon Suh for facilitating the survey of speech-language pathologists, and Oliver Niebuhr, Robert Xu, Alan Villamil, and Sara Pearsell for discussion. This work was supported in part by Otto Monsted's Fund and by the AI Research Institutes program of the NSF and the Institute of Education Sciences, U.S. Department of Education through Award # 2229873 – National AI Institute for Exceptional Education.

References

Andrea E. Abele and Bogdan Wojciszke. 2013. The Big Two in social judgment and behavior. *Social Psychology*, 44:61–62.

Nalini Ambady and Robert Rosenthal. 1992. Thin slices of expressive behavior as predictors of interpersonal

consequences: A meta-analysis. *Psychological Bulletin*, 111:256–274.

Naoki Azuma, Daichi Shikama, Asahi Ogushi, Toshiki Onishi, Ryo Ishii, and Akihiro Miyata. 2025. What timing and behavior patterns determine speed dating success in Japan? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*.

Radim Belohlavek and Tomas Mikula. 2024. Similarity metrics vs human judgment of similarity for binary data: Which is best to predict typicality? *Applied Soft Computing*, 153. Paper id: 111270.

Andrea Beltrama. 2020. Social meaning in semantics and pragmatics. *Language and Linguistics Compass*, 14(9):e12398.

Stefan Blohm and Mathias Barthel. 2025. Why so cold and distant? Effects of inter-turn gap durations on observers' attributions of interpersonal stance. In *Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue (SemDial, Bialogue)*, pages 125–134.

Laetitia Bruckert, Patricia Bestelmeyer, Marianne Latinus, Julien Rouger, Ian Charest, Guillaume A Rousselet, Hideki Kawahara, and Pascal Belin. 2010. Vocal attractiveness increases by averaging. *Current Biology*, 20(2):116–120.

Dana R Carney, C Randall Colvin, and Judith A Hall. 2007. A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, 41(5):1054–1072.

Sandeep Nallan Chakravarthula, Brian RW Baucom, Shrikanth Narayanan, and Panayiotis Georgiou. 2021. An analysis of observation length requirements for machine understanding of human behaviors from spoken language. *Computer Speech & Language*, 66. Paper id: 101162.

Herbert H Clark. 1996. *Using Language*. Cambridge University Press.

Elizabeth Couper-Kuhlen and Margret Selting. 2018. *Interactional Linguistics*. Cambridge University Press.

Neeraj Deshmukh, Aravind Ganapathiraju, Andi Gleeson, Jonathan Hamaker, and Joseph Picone. 1998. Resegmentation of Switchboard. In *ICSLP*, pages 1543–1546.

Maureen Fontaine, Scott A Love, and Marianne Latinus. 2017. Familiarity and voice representation: From acoustic-based representation to voice averages. *Frontiers in Psychology*, 8. Article 1180.

Jonathan B Freeman and Chujun Lin. 2025. A high-dimensional model of social impressions. *Trends in Cognitive Sciences*, 9:790–801.

Laura Fernández Gallardo and Benjamin Weiss. 2018. The Nautilus speaker characterization corpus: Speech recordings and labels of speaker characteristics and voice descriptions. In *Conference on Language Resources and Evaluation (LREC)*.

Jeroen Geertzen. 2015. Exploring age-related conversational interaction. In *SemDial*, pages 42–47.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of ICASSP*, pages 517–520.

John Grothendieck, Allen L. Gorin, and Nash M. Borges. 2011. Social correlates of turn-taking style. *Computer Speech and Language*, 25:789–801.

Judith A Hall, Summer E Harvey, Kirsten E Johnson, and C Randall Colvin. 2021. Thin-slice accuracy for judging Big Five traits from personal narratives. *Personality and Individual Differences*, 171. Paper id: 110392.

Judith A Hall, Phil Verghis, William Stockton, and Jin X Goh. 2014. It takes just 120 seconds: Predicting satisfaction in technical support calls. *Psychology & Marketing*, 31(7):500–508.

Nicole Holliday. 2023. Siri, you've changed! Acoustic properties and racialized judgments of voice assistants. *Frontiers in Communication*, 8. Paper id:1116955.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Pia V Ingold, Anna Luca Heimann, and Simon M Breil. 2024. Any slice is predictive? On the consistency of impressions from the beginning, middle, and end of assessment center exercises and their relation to performance. *Industrial and Organizational Psychology*, 17(2):192–205.

Hankyung Kim, Dong Yoon Koh, Gaeun Lee, Jung-Mi Park, and Youn-kyung Lim. 2019. Designing personalities of conversational agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*.

Inge S Klatte, Vera Van Heugten, Rob Zwitserlood, and Ellen Gerrits. 2022. Language sample analysis in clinical practice: Speech-language pathologists' barriers, facilitators, and needs. *Language, Speech, and Hearing Services in Schools*, 53(1):1–16.

Uus Knops and Anton M Hagen. 1989. Paradigms in sociophonetic research. In *New Methods in Dialectology*, pages 87–105. Foris Publications.

Michael W Kraus. 2017. Voice-only communication enhances empathic accuracy. *American Psychologist*, 72(7):644–654.

Nadine Lavan, Paula Rinke, and Mathias Scharinger. 2024. The time course of person perception from voices in the brain. *Proceedings of the National Academy of Sciences*, 121(26).

Hung Le, Sixia Li, Candy Olivia Mawalim, Hung-Hsuan Huang, Chee Wee Leong, and Shogo Okada. 2024. Do we need to watch it all? Efficient job interview video processing with differentiable masking. In *Proceedings of the 26th International Conference on Multimodal Interaction*, pages 565–574.

Gaël Le Mens, Balázs Kovács, Michael T Hannan, and Guillem Pros. 2023. Using machine learning to uncover the semantics of concepts: How well do typicality measures extracted from a BERT text classifier match human judgments of genre typicality? *Sociological Science*, 10(3):82–117.

Wenchao Li, Zhentao Zhong, and Haitao Liu. 2024. A computer-assisted tool for automatically measuring non-native Japanese oral proficiency. *Computer Assisted Language Learning*, pages 1–57.

Kevin D Lilley, Ellen Dossey, Michelle Cohn, Cynthia G Clopper, Laura Wagner, and Georgia Zellou. 2025. Social evaluation of text-to-speech voices by adults and children. *Speech Communication*, 166. Paper id: 103163.

Sabrina López, Pablo Riera, María Florencia Assaneo, Manuel Eguía, Mariano Sigman, and Marcos A Trevisan. 2013. Vocal caricatures reveal signatures of speaker identity. *Scientific Reports*, 3(1):3407.

Ioanna Lykourentzou, Robert E Kraut, and Steven P Dow. 2017. Team dating leads to better online ad hoc collaborations. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 2330–2343.

Magdalena Matyjek, Isabel Dziobek, Antonia Hamilton, and Thalia Wheatley. 2026. Social interaction style in autism: A critical review of social behaviours and outcomes in autistic and neurotypical interactions. *Autism in Adulthood*.

Phil McAleer, Alexander Todorov, and Pascal Belin. 2014. How do you say 'hello'? Personality impressions from brief novel voices. *PloS one*, 9(3). Paper id: e90779.

Katherine Metcalf, Barry-John Theobald, Garrett Weinberg, Robert Lee, Ing-Marie Jonsson, Russ Webb, and Nicholas Apostoloff. 2019. Mirroring to build trust in digital assistants. In *Interspeech*, pages 4000–4004.

Juliana Miehle, Isabel Feustel, Julia Hornauer, Wolfgang Minker, and Stefan Ultes. 2020. Estimating user communication styles for spoken dialogue systems. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 540–548.

Juliana Miehle, Wolfgang Minker, and Stefan Ultes. 2022. When to say what and how: Adapting the elaborateness and indirectness of spoken dialogue systems. *Dialogue & Discourse*, 13(1):1–40.

Nora A Murphy and Judith A Hall. 2021. Capturing behavior in small doses: A review of comparative research in evaluating thin slices for behavioral measurement. *Frontiers in Psychology*, 12. Paper id: 667326.

Laurent Son Nguyen and Daniel Gatica-Perez. 2015. I would hire you in a minute: Thin slices of nonverbal behavior in job interviews. In *ACM International Conference on Multimodal Interaction (ICMI)*, pages 51–58.

Shogo Okada, Yoshihiko Ohtake, Yukiko I Nakano, Yuki Hayashi, Hung-Hsuan Huang, Yutaka Takase, and Katsumi Nitta. 2016. Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 169–176.

Pilar Oplustil Gallegos, Jennifer Williams, Joanna Rownicka, and Simon King. 2020. An unsupervised method to select a speaker subset from large multi-speaker speech synthesis datasets. In *Interspeech*, pages 1758–1762.

Anastasia K Ostrowski, Jenny Fu, Vasiliki Zygouras, Hae Won Park, and Cynthia Breazeal. 2022. Speed dating with voice user interfaces: Understanding how families interact and perceive voice user interfaces in a group setting. *Frontiers in Robotics and AI*, 8. Paper id: 730992.

Dustin Palea, Ana Guo, Atira Nair, Ryan Anderson, Nisha Charagulla, Norman Makoto Su, and David T. Lee. 2024. Forming shared interest pods: Barriers to self-assembly of interest-based small groups, and dynamics of retaining and giving up control to find collective fit. *Proceedings of the ACM on Human-Computer Interaction*, 8.

Alex Pentland. 2007. Social signal processing: Exploratory DSP. *IEEE Signal Processing Magazine*, 24(4):108–111.

Rajesh Ranganath, Dan Jurafsky, and Dan McFarland. 2013. Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech and Language*, 27:89–115.

Felicia Roberts and Alexander L. Francis. 2013. Identifying a temporal threshold of tolerance for silent gaps after requests. *Journal of the Acoustical Society of America*, 133:471–477.

Gaetano Rossiello, Pierpaolo Basile, and Giovanni Semeraro. 2017. Centroid-based text summarization through compositionality of word embeddings. In *Proceedings of the MultiLing Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, pages 12–21. Association for Computational Linguistics.

Anke M Scheeren, Hans M Koot, and Sander Begeer. 2012. Social interaction style of children and adolescents with high-functioning autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 42(10):2046–2055.

Tanja Schultz. 2007. Speaker characteristics. In C. Müller, editor, *Speaker Classification I: Fundamentals, Features, and Methods*, pages 47–74. Springer.

Margret Selting. 1989. Speech styles in conversation as an interactive achievement. In Leo Hickey, editor, *The Pragmatics of Style*, pages 106–132. Routledge.

Tonya M Shoda. 2011. Interpersonal sensitivity and self-construals: Who's better at thin-slicing and when? Master's thesis, Miami University.

Atli Sigurgeirsson and Simon King. 2024. Controllable speaking styles using a large language model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10851–10855.

Lujun Su, Liqin Gong, and Yinghua Huang. 2025. Humorous or serious? The interaction effect of language style and tourism activity type on tourist well-being. *Journal of Sustainable Tourism*, 33:520–547.

Deborah Tannen. 1980. The parameters of conversational style. In *18th Annual Meeting of the Association for Computational Linguistics*, pages 39–40.

Deborah Tannen. 1990. *You Just Don't Understand: Men and Women in Conversation*. William Morrow.

Pol van Rijn, Silvan Mertes, Kathrin Janowski, Katharina Weitz, Nori Jacoby, and Elisabeth André. 2024. Giving robots a voice: Human-in-the-loop voice creation and open-ended labeling. In *CHI Conference on Human Factors in Computing Systems*.

Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. 2011. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87.

Sarah Theres Völkel, Ramona Schödel, Daniel Buschek, Clemens Stachl, Verena Winterhalter, Markus Bühner, and Heinrich Hussmann. 2020. Developing a personality model for speech-based conversational agents using the psycholexical approach. In *CHI Conference on Human Factors in Computing Systems*.

Michael Z Wang, Katrina Chen, and Judith A Hall. 2021. Predictive validity of thin slices of verbal and nonverbal behaviors: Comparison of slice lengths and rating methodologies. *Journal of Nonverbal Behavior*, 45:53–66.

Shuai Wang, Zhengyang Chen, Bing Han, Hongji Wang, Chengdong Liang, Binbin Zhang, Xu Xiang, Wen Ding, Johan Rohdin, Anna Silnova, and 1 others. 2024. Advancing speaker embedding learning: Wespeaker toolkit for research and production. *Speech Communication*, 162.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*.

Nigel G Ward and Jonathan E Avlia. 2023. A dimensional model of interaction style variation in spoken dialog. *Speech Communication*, 149:47–62.

Nigel G Ward, Andres Segura, Georgina Bugarini, Heike Lehnert-LeHouillier, Dancheng Liu, Jinjun Xiong, and Olac Fuentes. 2025. Towards precision characterization of communication disorders using models of perceived pragmatic similarity. In *IEEE ICASSP*.

Nigel G Ward, Andres Segura, Alejandro Ceballos, and Divette Marco. 2024. Towards a general-purpose model of perceived pragmatic similarity. In *Interspeech*.

Kwong-Cheong Wong and Jonathan Ginzburg. 2018. Conversational types: A topological perspective. In *21st SemDial*, pages 156–166.

Huili Yan, Tian Tian, Yifan Lu, and Hao Xiong. 2026. AI tour guides: The role of conversation style and destination type. *Tourism Review*, 81:755–781.

Ruiqing Zhao, Yubin Xie, Zichao Nie, and Ronggang Zhou. 2026. How do people communicate with conversational agents? Identifying communication styles through lexical and acoustic features. *International Journal of Human–Computer Interaction*.

A Appendix: More About the Experiment

This appendix provides more details on the experiment. The judgments and scripts are at https://www.cs.utep.edu/nigel/typical/.

A.1 Data

Our main reason for choosing to work with audio data alone was that this is more realistic for most of our use case scenarios.

Our main reason for choosing to work with utterances, rather than larger or smaller slices, was also because this seemed appropriate for our use cases. We chose to set a minimum utterance duration. In pilot studies, utterances of 5-seconds or more, unlike many shorter utterances, generally enabled listeners to form impressions about many speaker properties. For the experiments, we chose

to take only the first 5 seconds from every utterance, so that slice duration would not be a confounding factor for the comparisons.

We chose to examine the typical behavior of a speaker across their interactions with multiple partners. While previous thin-slicing work has focused on slices as representing a single, longer interaction, our use cases involve predicting something about the speaker in interaction with a new partner, so we chose to use information sampled from interactions with multiple others. This, however, raises the issue of style stability. Some aspects of style may be stable across partners, and for the others, a broad sample should ensure that less-common behaviors do not greatly affect the average.

To illustrate the nature of these conversations and clips, here are four examples of utterances that participants rated highly typical for a speaker:

> *uh, a landmark pool, it's, uh, a very excellent design* (Switchboard sw2422@1:37)
>
> *personally I guess my, my initial feeling is that, it, it would be a, a great* (sw2272@0:08)
>
> *what I was doing at, at home (in fact, I work nights here, and so that's another long story, we'll talk about)* (sw2629@0:03,)
>
> *I know, you can't find what, you used to, could find some for, you know, seventy-five dollars, a hundred dollars a month* (sw3003@4:38)

Incidentally, while these look very disfluent when transcribed, they sound natural and normal when listening.

As control conditions, we included three utterances selected without regards to our typicality model. One was randomly chosen, as the baseline. Another was the one closest to the middle of one of the conversations. Finally there was a "first," that is, an utterance that was closest to the beginning of one of the conversations, although, given the nature of the Switchboard recordings and the requirement for at least five seconds duration, this was never actually the first utterance of a conversation. Nevertheless we expected "first" to be a strong contender, because early utterances may be "purer" expressions of the speaker's true nature, since being less affected by the influences of the interlocutor and the vagaries of topic development.

As a technical detail, since the number of utterances was finite, we needed the selection algorithm to avoid duplicates. To do so, we chose the five utterances in order: first we chose the utterance closest to the centroid S. Then we chose a caricature utterance. If the one closest to the caricature point C was the same as the one already chosen, we picked the second closest. Similarly we then chose the first, the middle, and a random utterance.

A.2 Instruments

To get direct estimates of the likely utility of having typical utterances (RQ2), we asked the judges to rate their value for five scenarios. Scenario 1 represents our first use case, adaptation to an interaction partner, and Scenarios 2a-2d represent the second, choice of an interaction partner, with the specifics chosen to variously evoke situations where the interlocutor is mostly talking, mostly listening, or doing a lot of both.

As a side note, we did consider obtaining these utility ratings separately for each of the candidate typical utterances, but pilot studies showed that judges found it hard to do this. Instead, we asked these questions after they had heard five candidates for typical utterances and had chosen the best.

While direct ratings of typicality directly measure what we most care about, they have little ecological validity. We therefore included the descriptor-choice phases, representing a very common real-world activity: describing other people. We expected that the descriptions based on hearing a truly typical utterance would align with the descriptions after hearing a substantial behavior sample.

The judges did this first based on hearing each putative typical utterance: they choose descriptors to describe the speaker as he or she appears in that utterance. The prompt was:

- *Please provide adjective(s) to describe each utterance in terms of what it reveals about the speaker's conversational style or personality.*
- *Please use some adjectives from the following list, and add a few of your own.*
- *Adjectives:*
  - *Positive: Agreeable, Friendly, Pleasant (to listen to), Mature, Confident, Extroverted, Conscientious, Open-minded*

Value for	Rating	st.d.
1. If you knew you were going to meet this person, knowing in advance what sort of speaking style or interaction style you might use to connect well with them	3.5	±1.1
2a. Deciding whether you'd choose this person as your guide to a new city	3.0	±1.2
2b. Deciding whether you'd like this person to be your counselor, as a sensitive and supportive listener	4.0	±1.0
2c. Deciding if you'd choose this person to be on your team	3.7	±1.0
2d: Deciding if this is someone you wouldn't mind having a 10-minute conversation with	3.6	±1.1

Table 3: Utility Ratings, on a scale from 1 (no value at all) to 5 (very valuable)

  - *Negative: Disagreeable, Unfriendly, Irritating (to listen to), Immature, Insecure, Introverted, Careless, Closed-minded*

These suggested adjectives were selected from among the Big 5 personality traits and the descriptors from (Gallardo and Weiss, 2018).

Then, after participants listened to more data, specifically a minute of the speaker in a conversation, we solicited descriptors again:

> *Please provide adjective(s) to describe the speaker as he/she appears in conversation #1.*

and repeated this for one minute each from two more conversations.

We expected, relating to RQ 3 and RQ 4, that the descriptors chosen based on the caricature utterances would have a good amount of overlap with the descriptors chosen after the longer listening, and in particular greater than the overlap found for randomly chosen utterances.

Incidentally, we note that our decision to have participants choose their own descriptors departs from common practice in thin-slicing research, where people are given one or more attributes to judge. Using a small, fixed list of attributes makes it easy to compare the results of the thin-slice judgments to the full-session judgments, using simple measures such as correlation. However our choice aligns with what people would likely care about in our use cases.

A.3 Procedure

We chose to limit the set of candidates to only 5 utterances, to avoid tiring the judges.

We combined the methods into a single, four-phase protocol, because all three methods require participants to listen to a substantial amount of speech from each speaker, in Phase 2. Since this dominates the time cost, it was efficient to do this once per speaker and share that cost across all the judgments. While this is entirely appropriate for the first and second instruments, it is not so for the third, since it makes the descriptor choices not independent. In theory, ideal judges could perhaps have ascribed speaker descriptors based on each sample independently, without being swayed by the descriptors they had just given to a previous sample from the same speaker, but our real judges did not attempt this, as discussed below. Thus the actual value of obtaining descriptors was only for keeping the judges engaged and perceptive during Phases 1 and 2 and for post-hoc qualitative analyses.

A.4 Participants

All participants had connections with other participants, as friends, co-workers, and/or relatives, which likely motivated them to perform well. During the experiments, all sat around a large table, so they could see that everyone else was also paying close attention.

A.5 Results

We expected the descriptor overlap to be greater for the caricature condition, but there was no such effect. (Specifically, we ran a Poisson Generalized Linear Mixed Model (GLMM). Because the dependent variable, the number of overlapping adjectives, is a count, we used the Poisson family of distributions. The model had the condition as a fixed effect, and the judges ($n = 10$), speakers ($n = 11$), and the order of the utterances per speaker ($n = 5$) as the random effects, yielding the equation $\#DescriptorOverlap \sim Condition + (1|JudgeID) + (1|SpeakerID) + (1|OrderID)$. as the GLMM.) In fact none of the differences among conditions were significant, likely because

the overlap counts are all very low. There are two reasons for this. The first reason is that, with an unbounded set of descriptors, terms close in meaning but with slightly different nuances did not contribute to the overlap counts. The second reason is that, as hinted earlier, the judges apparently perceived their task to be, not making a sequence of independent judgments, but cumulatively discoving more facets of the person. We saw this in the discussion, and also in two aspects of their behavior. First, while we offered them the option to flag descriptors from the single-utterance case that they felt were not truly applicable, in light of subsequently listening to the three minutes of conversation, they never took this option. Second, they seldom bothered to repeat descriptors ascribed from the single-utterance case when picking descriptors for the 1-minute samples. Overall we do not regret our choice to allow our judges to choose descriptors freely, but it did thwart our desire for an additional way to show statistical significance.

Table 3 shows the utility ratings. As a secondary effect, we observed that the utility ratings varied with the speaker being judged, and were unusually low for one speaker. Listening, we found that all the clips for her were instances of cheerfully telling stories, which were apparently felt by the judges to provide little insight into her likely behavior in the various scenarios.