Interaction Styles

A Dimensional Model of Interaction Style Variation in Spoken Dialog

Nigel G. Ward*  Jonathan E. Avila**
University of Texas at El Paso  University of Texas at El Paso

In spoken dialog people vary their interaction styles, and dialog systems should also be able to do the same. Previous work has elucidated many aspects of style variation and adaptation, but a general model has been lacking. We here describe a dimensional model of the space of interaction styles, derived by applying Principal Component Analysis to 33022 conversation fragments from the Switchboard corpus of American English telephone speech, represented by 84 novel features that encode the frequencies of diverse interaction-related prosodic behaviors. The top 8 dimensions were meaningfully interpretable, and include aspects previously noted in the literature but also new ones. Both this vector space representation and the method used to derive it may be useful for dialog systems design, tuning, and adaptation. Further, regarding individual differences in interaction style, we find that individual style tendencies were surprisingly weak, with a predictive model based on individual tendencies outperforming a speaker-independent model by only 3.6%.

1. Introduction

Interaction styles vary among people and across situations. The importance of these styles is seen in our rich folk vocabulary for describing them, including terms like agreeable, quick, playful, chatty, reserved, soft-spoken, serious, domineering, and many more. Many aspects of interaction style relate to wider social issues and even lifelong patterns of behavior, as addressed by work on conversational styles and communication styles, but our focus here is on styles that people use transiently in the course of dialog. While previous research has diversely identified a number of interaction style aspects, there are many practical and scientific reasons to want a consistent general model of the space of variation in interaction styles.

This paper describes how we used Principal Component Analysis over a large set of features that represent behaviors in spoken dialog to derive a space of interaction styles. We describe the 8 top dimensions, including several not noted in previous work. Further, we investigate individual differences, and find, that, while most individuals exhibited interaction style tendencies, these were generally far from stable. We also examined gender differences in style and adaptation, among other related questions.

* E-mail: nigelward@acm.org
** E-mail: jonathan.edav@gmail.com

© 2021 Association for Computational Linguistics
2. Motivations

Lack of a general model of interaction style variation has impeded progress towards many important goals:

2.1 Making Dialog Systems more Adaptive

We would like our dialog systems to be able to work well for any user. User modeling and adaptation are venerable topics in interactive systems research, with most work on style in dialog addressing style choices in specific components, such as speech synthesis, utterance selection, and language generation (Eskenazi 1993; Wang et al. 2018; Cao et al. 2020; Hu, Fox Tree, and Walker 2018; Niu and Bansal 2018; Cheng, Fang, and Ostendorf 2019; See et al. 2019; Chaves and Gerosa 2020; Gordon et al. 2021; Jin et al. 2020). However, we ultimately would like our systems to be able to adapt their overall interaction style: after inferring the user’s style, and preferences regarding style, a dialog system should adjust its behavior to suit, in order to improve efficiency and acceptability for each specific user (Eskenazi and Zhao 2020; Marge et al. in press, 2021). This goal is now becoming achievable, as in Metcalf’s adaptation of Siri to detect which users are more talkative and then provide them information in a more chatty style (Metcalf et al. 2019). While this system arose from a specific insight and involved careful engineering, a general model of the space of interaction styles could support more such adaptations. Moreover, one-by-one adaptation of individual elements of interaction style, without considering interactions among elements, risks creating dispreferred or inauthentic system behavior (Gordon et al. 2021).

Interaction style control will likely also become an issue for multifunctional dialog systems, such as those capable of both task-oriented and chitchat interaction (Sun et al. 2021), where the style may need to change to suit the current dialog activity.

Modeling interaction style variation could also indirectly support the development of better dialog systems, by enabling the creation of individualized preference models, that is, models of the mapping from system behavior to user satisfaction (Yang, Levow, and Meng 2012; Ward 2019a; Gordon et al. 2021), which could support more tailored adaptations, either offline or online.

2.2 Making Dialog Systems better Targeted

We would like to support more rapid development of dialog systems. Today, it remains common practice to gather a new corpus for every new development project, to support extraction of the various knowledge and parameters needed. Instead, we should be able to better leverage existing corpora (Kashyap et al. 2021). For example, Pecune et al. (Pecune et al. 2019) created a recommendation system by exploiting the subset of Switchboard data relating to movies. Selecting this subset was easy for them because Switchboard was designed around topics, and in particular the “movies” tag was available. However, in general, we would like to be able to rapidly delineate corpus subsets that match some specified target style, and a model of the space of interaction styles would enable us to pick out subsets that are close to some set of seeds.

Modeling interaction styles could also support more targeted corpus development. Several recent large efforts have aimed to collect dialogs exemplifying a sweet-spot style that is simultaneously natural for humans and implementable with current technology (Budzianowski et al. 2018; Byrne et al. 2019). However, for lack of a model of styles, coarse manipulations are common, such as having dialogs be written down and then
read, or interposing a text-to-speech engine between the confederate and the subjects. A reference model of styles would help us better identify how such manipulations affect style and how well the result matches an application need.

In the same vein, models of interaction style may ultimately allow style mixing, to create, say, a system whose style is 70% that of Taskmaster, with a 10% addition of the error-resolving style elements of Maptask, with a 15% addition of the role and personality elements of a proprietary restaurant server corpus, and with a 5% touch of the style elements of the corporate brand. Making such mixing possible will finally enable full exploitation of the rich variety of existing corpora.

2.3 Systematizing Computational Pragmatics

The field of pragmatics has so far been only sporadically cumulative: it is dismayingly common for reports of a correlation between some feature and some perception to not be confirmed by follow-on studies. This can be seen, for example, in work on turn taking, politeness strategies, rapport building, and quality perceptions. A better understanding of interaction styles may resolve such divergences. For example, studies of communications channels had variously found the effects of delay to be significantly disruptive or quite tolerable, but the contradiction was resolved when the dimension of "interactivity" was identified: conversations in more interactive styles are now known to be more sensitive to delay (Egger, Reichl, and Schoenenberg 2014). Other conflicting results, such as findings that women use creaky voice more, but also less, and that women tend to accommodate more, but also less, remain to be reconciled (Wright, Mansfield, and Panfili 2019). A likely cause of these confusions is the current inability to characterize corpora other than in terms of superficial properties, such as the demographics of the participants, the dialog goal, and the typical topics. Deeper models, describing how dialog sets vary in interaction styles, could enable us to properly map out the field of inquiry of computational pragmatics, extending the current collection of scattered observations across unrelated corpora into a unified field of study.

2.4 Supporting Improved Human-Human Communication

We would like to be able to help people communicate more successfully, especially when differences in interaction style can cause misattributions and misunderstandings (Tannen 1989, 1990; van Kleeck 1994). The need for awareness of such differences is widely felt, as attested by the many self-help books on overcoming culture-linked and gender-linked style differences. However, much of this advice is based on anecdotal observations, and better models of style variation could support the development of more evidence-based advice. This could be useful either for offline training or for online nudging, to help individuals or small groups adopt more effective interaction styles.

We would also like to be able to provide individualized advice, through the automated assessment of abilities and propensities in interaction. Today, it is common to use some single scalar measure of conversational competence to score people, for example those learning a second language, or those being screened for autism. However, we would likely find more informative assessments, that describe abilities across specific styles, to support selection of individually appropriate instructional materials or therapy. We therefore need ways to better understand the heterogeneity of social communication modes, skill sets, and difficulties (van Kleeck 1994; Hudry et al. 2013; Nelson et al. 2014).
Thus an improved understanding of interaction styles could lead to new ways to help both children and adults increase their interaction style repertoire and social effectiveness.

3. Previous Work

A good model of the space of interaction styles could have many benefits. Fortunately, previous research has done valuable groundwork. Although we here use the term “interaction styles,” to indicate our focus on the properties of interactions per se, most related work has used other terms — including conversational styles, interpersonal styles, style matching, and so on — reflecting variously a broader or narrower focus.

The earliest work on interaction styles was motivated by the observation that a major cause of social awkwardness can be style differences. Tannen and others noted how aspects of conversational style that are normal for one subculture can be perceived as cold or rude by members of another subculture. Other work has explored how gendered, age-related, and culture-specific aspects of interaction style can be misperceived (Tannen 1989, 1990; Georgila et al. 2010; Geertz 2015). This is an active and complex area of research, not least because aspects of style also relate to genre (Biber 2004; Prsir, Goldman, and Aucelin 2014), topic (Pavalanathan et al. 2017), stance taking (Kiesling 2009; Ranganath, Jurafsky, and McFarland 2013; Lai et al. 2019), personality (Burger 1990; Berens 2001; Yu, Gilmartin, and Litman 2019; Yamamoto et al. 2020), culture (Endrass, Rehm, and André 2009), social role, self presentation, and social identity (Tannen 1987).

This tradition has identified numerous dimensions of interaction style, of which the most frequently discussed include high-contact vs high-consideration, verbose vs concise and focus-on-content vs interpersonal involvement. Such dimensions have been influential in the design of interactive systems and virtual agents (Miehle et al. 2020), sometimes filtered through the lens of system or user personality (Bevacqua et al. 2012; Kuzminykh et al. 2020; Völkel et al. 2021), and are increasingly relevant in the customization and adaptation of spoken dialog systems (Cohen, Giangola, and Balogh 2004; Shamekhi et al. 2016; Lukin et al. 2018; Hoegen et al. 2019; Metcalf et al. 2019). While useful and quantifiable (Wei et al. 2014), these dimensions do not provide a systematic or general model of the space of interaction styles.

A second major motivation for modeling interaction styles originates in the observation that people in dialog are flexible, and adapt their behavior to their conversation partner (Giles et al. 1987). Dialog systems that fail to do so can seem unappealing and robotic. This has motivated attempts to identify how people vary and adapt to each other, studied variously under the rubrics of accommodation, entrainment and convergence. Work in these traditions tends to focus on simple, directly measurable properties of an individual’s speech, such as counts of different word classes, average pitch, and average articulatory precision (Niederhoffer and Pennebaker 2002; Levitan et al. 2012; Litman et al. 2016; Weise et al. 2019; Lubold et al. 2019; Galvez et al. 2020). Unfortunately, these findings do not reliably generalize across corpora (Levitan 2020), and again do not provide a systematic model.

Thus there is, on the one hand, a mostly qualitative tradition that looks at “macro” aspects of style, aligning with socially meaningful practices, each often involving multiple indicators, and, on the other hand, a quantitative tradition that looks at directly measurable, “micro” features of dialog. (Both of these traditions are also seen in social psychology work on group dynamics.) Yet there are commonalities in the phenomena
Ward & Avila Interaction Styles

studied by the two traditions. For example, overlapping speech can be treated as either a low-level feature or a macro-level style parameter.

Connecting these two traditions is a third, more recent, approach. This uses the correlation patterns among micro features to discover macro dimensions of variation. To briefly summarize the dimensions found by work in this vein, deferring discussion of methods until later: Biber (Biber 2004) found three dimensions: information-focused vs interactive discourse, stance-focused vs context-focused discourse, and narrative-focused vs present-tense discourse. Grothendieck et al. (Grothendieck, Gorin, and Borges 2011) identified two clusters, one involving more overlap, less silence, and shorter turns, which they tentatively identified as “female-style,” with the other, “male-style” cluster having the opposite characteristics. Laskowski found three dimensions of individual behavior, representing time spent talking, the inclination to overlap, and social status (Laskowski 2016).

These studies have shown that it is possible to connect the two levels of inquiry: confirming that perceptions of macro-level style dimensions can be rooted in specific, measurable behaviors. However, so far the coverage of these models has been limited.

The primary aim of this paper is, accordingly, to extend this approach to identify and quantify more dimensions of interaction style.

4. Methods

Our strategy follows that of Biber (Biber 2004): we apply Principal Component Analysis to suitable features over a large set of conversation data in order to derive a dimensional representation of the space of styles. Our data choice follows Laskowski and Grothendieck et al. (Laskowski 2016; Grothendieck, Gorin, and Borges 2011): we use spoken dialog data because it contains more indications than text. Our features are original, designed to capture a broad variety of interactive and prosodic behaviors. This section gives the specifics.

4.1 Model

We model each instance of an individual speaker’s conversation behavior as a datapoint in a space of styles. Although clustering approaches can have advantages (Biber et al. 2021), we here follow most work in this area, dating back to the earliest work by Tannen (Tannen 1980), in seeking a dimensional description, as this can represent more generalizations and can support more precise characterizations.

Among the many ways to derive such a space, we chose Principal Component Analysis (PCA). Each dimension output by PCA is a linear combination of the input features, without ambiguity or complexity, which greatly simplifies interpretation. Moreover, PCA-derived dimensions often turn out to be meaningful factors that explain the observed, surface variation. PCA is also robust to noise.

4.2 Features

To apply PCA, we need to compute suitable features for each datapoint. For the quantitative study of interaction styles, there are many choices for what features to use.

Biber used lexical features, namely a selection of discourse markers, parts of speech, and semantic classes, augmented with automatically-parsed grammatical structure elements (Biber 2004). However, at least for spoken conversation, such features have
so far revealed only three clear dimensions, far fewer than the number suggested by qualitative research.

Other work has used turn-taking features, often based on sequences of states, indicating at each frame which speaker is talking, or both, or neither (Yamamoto et al. 2020; Grothendieck, Gorin, and Borges 2011; Laskowski 2016). Features have included dwell times and transition probabilities. The advantage of such features is that they are strongly linked to the style of the interaction itself, and less influenced by confounds such as dialect and topic. However, previous research indicates that such features alone can reveal only a few dimensions of interaction style. Moreover, turn-taking does not operate in a vacuum; as will be seen, the turn-taking aspects of interaction styles are intimately related to other dimensions of variation.

In this work we chose instead to use prosody-based features. Prosody serves many purposes in conversation, and thus has the potential to reveal a richer picture. While one might combine prosodic and lexical features, here we derive the model using only prosody-based features, mostly because this allows us to use lexical features as an independent source of evidence for interpretation.

The rest of this section elaborates on the specific prosodic features used, and how they manage to represent many aspects of dialog behavior. Readers not interested in the details may profitably skip ahead to Section 5.

4.2.1 Feature Design Strategy. Having chosen to use features based on prosody, there are still many choices. While it may be someday possible to use a comprehensive feature set — one that reflects all possible behavior sequences and configurations, and their exact realizations, and their timings relative to each other and to the interlocutor’s behaviors — to avoid a combinatorial explosion, we here need to be selective.

It’s worth stepping back for a moment, to consider where interaction style differences are mostly likely to manifest themselves. First, it seems likely they will be most evident in short, several-second behavior sequences, rather than, on the one hand, transient events or, on the other hand, long-term average behaviors. For example, in peer-tutoring dialogs, there is a recurring behavior sequence in which one peer insults the other and then produces a glance and a smile (Zhao et al. 2016), and this temporal configuration of activities is a marker of a teasing/familiar interaction style. Second, it seems likely that interaction style differences will be most evident not in the use of rare patterns, but in the use of generally informative patterns.

Applying these considerations to the selection of prosodic features, they suggest a focus on temporal configurations of (often) joint behaviors, and the use of generally informative configurations. While finding such configurations can be a challenge in itself (Janssoone 2015), fortunately, for English prosody there is already a convenient inventory of the most common ones. The next subsection explains.

4.2.2 Prosodic Constructions. The features we use represent aspects of prosodic construction use. A prosodic construction is simply a temporal configuration of prosodic properties with a meaning or function (Ogden 2010; Niebuhr 2014; Ward 2019c).

A simple example is the Positive Assessment Construction, illustrated in Figure 1 (Ward and Jodoin 2019). Like most constructions, it involves multiple prosodic properties — pitch, rate, and intensity — in a specific temporal configuration. Although it is often aligned with two or more words, this construction, like most, can align quite flexibly with words and syntactic structures. Like all prosodic constructions, this prosodic form can be present to a greater or lesser degree, proportional to the strength of the feeling or function. Also like all prosodic constructions, it can serve a variety of
Backchanneling Construction

Functions: encourage continued listening; encourage continued talk, etc.

Form:

<table>
<thead>
<tr>
<th>timespan</th>
<th>prosodic properties, etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>loudness, creakiness, and articulatory precision increase</td>
<td></td>
</tr>
<tr>
<td>pitch drops and stays low</td>
<td></td>
</tr>
<tr>
<td>loudness decreases</td>
<td></td>
</tr>
<tr>
<td>silence</td>
<td></td>
</tr>
<tr>
<td>backchannel: lengthened, quiet, flat pitch</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2
The Backchanneling Construction.

related purposes, here including showing respect to the interlocutor, praising someone or something, and expressing admiration.

Another example is the Enthusiastic Overlap Construction. This construction is a joint behavior, produced by the actions of both speakers. It involves, in the ideal realization, both speakers speaking or laughing, with wide pitch range and a slightly reduced speaking rate, over about 500 milliseconds. This relates to turn-taking processes, but is not just a superficial measure of whether two voices are simultaneously active, which may happen for many reasons, including inadvertently. Rather, it describes a recurring turn-taking behavior that involves specific prosody and conveys a specific family of functions, mostly involving some kind of enthusiasm. Similarly the Turn Hold Construction involves a person not merely speaking, but also actively holding the floor with moderately high but slowly declining pitch, and rather slow speaking rate. There is also the Turn Exchange Construction, in which one speaker signals an upcoming end, and the other times their entry tightly and comes in with elevated speaking rate and high pitch.

Another example is the Backchanneling Construction, described in Figure 2. Functionally, use of this construction demonstrates one participant’s intent to continue in the speaking role and to produce the next installment of information, and the other participant’s interest and intention to continue listening. The timing and prosodic properties
in the figure represent an ideal realization of this pattern, but any specific occurrence will differ. For example, the listener may omit the backchannel, or produce one that is atypically early, late, quiet, or expressive; and the speaker may produce the backchannel cue weakly or strongly, or may deliver the second installment without leaving a gap for the backchannel, and so on. Some of these effects may be due to the effects of simultaneously-present constructions, for example, an enthusiastic backchannel — with higher pitch, wider pitch range, and overlapping the ongoing talk of the main speaker — may reflect the superimposed use of the Enthusiastic Overlap Construction.

The constructions used in this study are listed in Table 1, using short names that evoke the most common functions of each configuration. As described in (Ward 2019c), they were derived automatically, from data. The process started by computing 212 speaker-normalized prosodic measures, over windows spanning about 3 seconds, at hundreds of thousands of timepoints spaced every 20 milliseconds across 80 minutes of American English conversations. Next PCA was applied. (To clarify, this application of PCA is prior to and unrelated to that used to discover the dimensions of interaction style.) As features from both speakers were used, there was a bias to the discovery of behaviors involving both speakers (Ward and Jodoin 2019). However, unlike many descriptions of meaningful prosody, they are fully quantitative. Behind the concise names in the table, each construction is a complex configuration of multiple prosodic properties — including actions by both participants, and quantified by the loadings over the 212 measures (Ward 2019b) — and used for a context-dependent variety of related purposes.

Many of these constructions correspond to patterns well-known in the literature, and for some, the imputed pragmatic functions have already been confirmed by human-subjects experiments (Kurumada, Brown, and Tannenhau 2012; Day-O’Connell 2013; Ward and Jodoin 2019). However, unlike many descriptions of meaningful prosody, they are fully quantitative. Behind the concise names in the table, each construction is a complex configuration of multiple prosodic properties — including actions by both participants, and quantified by the loadings over the 212 measures (Ward 2019b) — and used for a context-dependent variety of related purposes.

<table>
<thead>
<tr>
<th>#</th>
<th>Construction</th>
<th>Other Construction</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>focal has turn</td>
<td>other has turn</td>
<td>Turn Hold</td>
</tr>
<tr>
<td>2</td>
<td>silence</td>
<td>enthusiastic overlap</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>turn take by focal</td>
<td>turn take by other</td>
<td>Turn Exchange</td>
</tr>
<tr>
<td>4</td>
<td>other backchanneling</td>
<td>focal backchanneling</td>
<td>Backchanneling</td>
</tr>
<tr>
<td>5</td>
<td>topic closing</td>
<td>topic continuation</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>topic development</td>
<td>positive assessment</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>empathy bid by focal</td>
<td>empathy bid other</td>
<td>Empathy Bid</td>
</tr>
<tr>
<td>8</td>
<td>turn-hold fillers</td>
<td>bipartite construction</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>long turn take</td>
<td>long turn yield</td>
<td>Long Turn Exchange</td>
</tr>
<tr>
<td>10</td>
<td>late peaks</td>
<td>turn-initial fillers</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>meta comment</td>
<td>bookended narrow pitch</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>minor third cue</td>
<td>response to an action cue</td>
<td>Minor Third</td>
</tr>
</tbody>
</table>

Table 1
The Prosodic Constructions Used in the Feature Computations
These constructions were generated from one specific corpus with one specific set of measures (Ward 2019c), but experiments with other corpora and other measures yield fairly similar inventories (Ward 2014; Ward and Gallardo 2017). Here we use this specific inventory primarily because it is the best documented.

Using prosodic behavior configurations as the basis for analysis has several advantages. Compared to raw prosodic measures, they have meaningful functions, and are therefore more likely to relate to meaningful style differences. They include a diversity of patterns, including joint behaviors where the participants are taking opposite but complementary roles, joint behaviors where the participants are doing similar things, and individual behaviors. Given their quantitative nature, the strength of their presence can be computed at every moment in time, fully automatically, with proper normalization to each speaker, using public-domain software (Ward 2021c). Finally, critically, they cover a wide range of dialog states, activities, and events, including many of those often considered most important in human interaction (Borger 2018; Couper-Kuhlen and Selting 2018), and they involve most of the low-level features examined in research on micro-level style aspects.

### 4.2.3 Features that Quantify Distributions over Prosodic Construction Uses.

Having decided to use features representing prosodic behaviors, and to base them on prosodic constructions, the next question was what exactly to quantify.

We want the features to represent the extent to which the speakers are engaged in various interaction routines, and also the extent to which the dialog tends to dwell in certain states. Statistics on the prosodic construction usage can serve both purposes. For example, the fraction of timepoints at which the Enthusiastic Overlap Construction is strongly present indicates the frequency of strong engagement, the fraction where it is weakly present indicates the frequency of mild engagement, and the fraction where there is no evidence for it indicates the prevalence of lack of engagement.

We accordingly define each feature to be the fraction of timepoints for which a given prosodic construction dimension value is in a certain range. The feature computation process is straightforward. First it computes the quality of the match between each prosodic construction dimension's configuration and the interactants' behavior, every 20 milliseconds across each conversation fragment. Next it computes the frequencies of occurrence for each range. Outliers, times where a dimension's value is further than 10 standard deviations from the norm, are ignored; the rest are counted in seven bins. There is one bin for the fraction between −0.4 and 0.4 standard deviations, representing the evidence for the absence of either construction of a pair, and three bins on each side of the mean, namely more than 2.4 standard deviations away, 1.4 to 2.4 away, and 0.4 to 1.4 away. Thus each conversation fragment is characterized by 84 bin frequency features (BFFs): 7 bin frequencies each for the 12 prosodic construction dimensions named in Table 1.

Figure 3 overviews this process. This method effectively aggregates prosodic construction occurrence information across the dialog fragments we wish to characterize. Despite occasional inaccuracies due to noise or non-modeled aspects of prosodic behavior, the resulting features are fairly robust. The computation is fully automatic, and the features are fairly easy to interpret. They thus provide a good number of meaningful features that cover many important interactive behaviors and states.
Figure 3
Illustration of the Feature Computation. The input audio signal, left and right tracks, is shown in the top two plots. The next three plots show the values of prosodic principal components 1 through 3. The numbers at right show the fraction of time spent in each of the seven ranges. We see that the Dimension 1 features correctly capture the fact that neither speaker ever really held the floor (the .00s), but that Speaker A more often approached this role (the .34). The Dimension 2 features capture the fact that there are unusually large amounts both of overlap and of silence in this fragment. The Dimension 3 features accurately represent the prevalence of low-overhead/low-commitment rapid turn-taking. The audio file used to generate this figure is available at http://www.cs.utep.edu/nigel/istyles/.
4.3 Data and Processing

For analysis we chose the Switchboard corpus of American English telephone conversations (Godfrey, Holliman, and McDaniel 1992; ISIP 2003), mostly for its diversity and size. Excluding dialogs with poor audio quality, and holding out some data for future studies, we chose a subset of 1426 conversations for building the model (Ward 2021b). As interaction styles are neither instantaneous nor stable over long times, we chose 30-second fragments as the unit of analysis. Thus each conversation was split into 30-second fragments, with any remainder discarded. With most conversations 5 or 10 minutes long, this gave 16511 fragments.

Each fragment was double-sampled, once treating the A (left) speaker as the “focal” speaker and B as the “nonfocal speaker,” then again with the roles reversed. This yielded $2 \times 16511 = 33022$ data points. This is an order of magnitude larger than in any previous attempt to model interaction styles (Biber 2004; Grothendieck, Gorin, and Borges 2011; Laskowski 2016).

Principal Component Analysis interpretation feature computation …
…
each fragment represented by 84 features 84 principal components, each a loading over the 84 features

dimensional representation of the style of fragment X

Figure 4
Overview of model construction and use.

The feature computation described above was applied to each fragment, then the features were z-normalized and PCA was applied, as seen across the top of Figure 4.

Code for the entire process is available at (Ward 2021b). This can characterize new data using the existing BFFs, as seen at the bottom of Figure 4, or, for analysis of languages other than English, can derive new BFFs, fully automatically, from data.

4.4 The Interpretation Process

The result of PCA is, by mathematical necessity, a dimensional representation of the space of variation. While this may be useful as-is for some purposes, our primary aim is to extend our understanding of the space of interaction styles, so we sought to use these dimensions to discover meaningful aspects of style.

As always when trying to characterize an inferred latent structure, one can consider the way the structure was derived or examine the datapoints that it characterizes in one way or another. Here we do both. Specifically, as each dimension has two poles, positive and negative, we looked at each in turn, considering evidence of three kinds.
The first kind of evidence is the loadings of the dimension on the features.

The second kind of evidence is lexical tendencies. Conversation fragments towards a pole are often rich in occurrences of words of one kind or another. To quantify this, for the negative pole we counted the occurrences of all words in all conversation fragments below the $10^{th}$ percentile on the dimension, and for the positive pole, fragments above the $90^{th}$ percentile. Limiting attention to words occurring at least 30 times in the corpus, we computed which were more frequent in the near-pole fragments than in Switchboard overall, and scanned the listings of those whose frequencies across those fragments were less than half or more than double the base frequency, and in addition the frequency ratios for the 40 most frequent words. We also examined word-class tendencies using LIWC (Pennebaker et al. 2015).

The third kind of evidence is subjective impressions of a sampling of the conversation fragments towards each pole. For the negative pole, we examined mostly fragments below the $3^{rd}$ percentile, and for the positive pole mostly above the $97^{th}$ percentile.

Considering all the evidence, we looked for commonalities and used these to form interpretations that would account for most of the evidence.

While we hoped that the interpretation phase would be easy, in fact it often took some effort to arrive at a satisfactory interpretation for a dimension. For some poles, some types of evidence were scant or difficult to interpret, so we had to rely more on the other types. Although examination of the fragments was reliably informative, for most poles there were one or more fragments which did not seem to align with any possible general interpretation. Sometimes these reflected clear confounds, such as an extended silence while one participant went to answer the door, and some were due to clear countervailing factors, but some simply seemed not to fit the general pattern. Another problem was the huge number of lexical tendencies for each pole, most of which did not fit into any obvious pattern. When we did find a pattern explaining a significant fraction of words, we checked also for counterevidence by examining the frequency of related words. For example, politics-related words, including Republican, but not Democrat, were relatively frequent at one pole of one dimension, so we looked at the words common at the other pole, to see whether the word Democrat was more frequent there. However, in no case was there such counterevidence. In general we did due diligence on all the evidence reported below, however, in the interests of space, we omit those details and report only selected evidence.

Finally, as a plausibility cross-check, after arriving at the interpretations, we examined topic tendencies. Every Switchboard conversation had a suggested topic. We considered topics whose distributions on some dimension diverged from the overall distribution ($p < 1e-12$, two-tailed t-test, using a very strict threshold to select only the most informative topics). Unlike the other kinds of evidence, for topic tendencies below we report all that met this criterion. Such topics were in general easy to understand as relevant to the interpretation of that pole.

Each dimension is richly complex, but the descriptions below are concise, describing only some of the most informative evidence, in order to give the reader a feel for the nature of each pole and dimension. More information, including the evidence of all four kinds for each pole, is available at the companion website (Ward 2021b).

5. Eight Dimensions of Interaction Style

This section describes 8 dimensions of interaction style.
5.1 Dimension 1, Positive Pole

The positive pole of Dimension 1 represents a lack of shared engagement.

The first source of evidence is the loadings, seen in Table 2. Features with positive loadings include those for silence (marked a), and the central ranges for every kind of turn taking and backchanneling (b). Thus this pole involves a lot of silence and a dearth of the usual turn-taking patterns.

The second source of evidence is the lexical tendencies: common words characteristic of this pole include um, uh, think and that. In terms of the LIWC word categories, this pole is strong on “work,” “money,” and “analytic.”

The third source of evidence is subjective impressions: listening to fragments near this pole, often these are times when the participants seem tired of a topic or the conversation itself, and seem to be continuing just to run out the clock.

The topics significantly high on Dimension 1 were “Soviet Union,” “capital punishment,” and “public education,” all topics for which the prompt suggested a formal dialog, for example with “take an opposing view in your discussion.”

Table 2
Feature Loadings of Interaction Style Dimension 1. The labels for the degenerate dimensions refer to the behavior of the focal speaker unless otherwise specified.

5.2 Dimension 1, Negative Pole

The negative pole of Dimension 1 represents a style of both participants engaged.

For this pole, the loadings are of course just the opposite of those for the positive pole. Accordingly, from the negative numbers in Table 2, we see tendencies to enthusiastic overlap (c), and every kind of turn taking (d), suggesting that this pole involves engagement and interactivity.

Common words characteristic of these fragments include [laughter], and oh. Other relatively frequent words include place names and hi and hello, both over 4 times as common as normal in these fragments. The LIWC categories show tendencies to “positive tone” and “positive emotion.”

Dialog activities common in fragments near this pole include learning interesting things about the nonfocal participant, such as where they live, their occupation and hobbies, and how they joined the corpus collection effort. More subjectively, in frag-
ments near this pole, the speakers seem interested and engaged, often clearly enjoying the topic and the conversation.

The topics significantly low on Dimension 1 were “football,” “baseball,” “TV programs,” and “movies,” all topics which people generally have a lot of experience with in light conversation.

Thus Dimension 1 represents a continuum between low and high engagement. Not only are the loadings opposite (a mathematical necessity), but the semantics of the two poles also turn out to be opposite. Perhaps surprisingly, this turns out to be largely true also for the next seven dimensions, as we will see.

5.3 Dimension 2, Positive Pole

Across Dimension 1 the two speakers tended to behave similarly, either both engaged or both unengaged, but in Dimension 2 they take different roles. To reduce confusion we will refer to them as the “focal” speaker and the “nonfocal” speaker; of course these are not fixed roles, and each individual contributor in the corpus is considered in each role, through the double sampling.

The positive pole of Dimension 2 seems to involve the focal speaker listening actively, and the nonfocal speaker mostly talking. The evidence is, again, diverse and consistent.

The prosodic loadings, seen in Table 3, indicate that here the nonfocal speaker mostly has the turn (e). The speakers frequently perform a weak turn exchange, with the nonfocal speaker yielding to the focal speaker (f). The focal speaker also tends to be prolifically backchanneling (g). Evidently at this pole the focal speaker is in a supporting role, with the nonfocal speaker doing most of the talking. The nonfocal speaker also has a tendency to use the configuration involving high pitch, narrow range, and high articulatory precision (h), which is generally a mark of speaking with indifference, in contrast to speaking to gain empathy or agreement.

Examining the lexical tendencies, for the focal speaker, uh-huh and uh-hum are both over 10 times as common in these fragments as elsewhere. wow, jeez, whoa, right, yep, yes and [laughter] are also very common, and the LIWC category of positive emotion is very strongly present.

Listening to some examples of this pole, they tend to involve long exposition by the nonfocal speaker — for example, regarding local manufacturing, German perceptions of America, or a local criminal case — with the focal speaker actively listening. In particular, the focal speaker’s weak turn taking actions (f) are often briefly supportive statements, such as sure, exactly, and oh, I felt that way too.

---

Table 3
Feature Loadings of Interaction Style Dimension 2. Rows not shown, in this and subsequent tables, have negligible loadings for all features (all with absolute values < 0.01).
5.4 Dimension 2, Negative Pole

The negative pole of Dimension 2 is the opposite of the positive side, and thus has focal speaker mostly talking, with the nonfocal speaker being an active listener. Words characteristic of the focal speaker for this pole include and, was, the, and low-frequency content words, such as wallpaper, cent, behavior, Republican and accounting. The LIWC “past focus” category is strongly present.

While no topics were significantly high or low on Dimension 2, the squared distance from the mean was relatively high for “consumer goods”, for which callers were prompted to talk about a recent “return a product bought recently” experience, and often did so in a long story. The squared distance was low for “vacation spots” and “music,” likely because speakers on these topics tended to exchange shorter observations, comments, and questions.

5.5 Dimension 3, Positive Pole

This pole relates to a negative feelings stance.

This is clear from the lexical tendencies, where, for example, gang, gangs, convicted, stole, offense, and disagree all occur over 3 times more commonly in fragments near this pole. This is also clear from the topics in fragments near this pole, which include include income tax, lawn problems, the futility of overseas aid, and time flying by. While a pole relating to negative feelings as a style surprised us, as we usually think of negative feeling as an individual attitude or stance, in this data there is clearly a recurring pattern of both speakers feeling negatively about something and, through the dialog, reinforcing each others’ opinion.

The significance of prosodic evidence for this dimension (Table 4) was not clear to us at first, but became clear after listening to representative fragments. Thus, instead of prosody helping us interpret the dimension, the existence of the interpretation enabled us to discover something about the prosodic behaviors associated with negative stance in English. This pole was associated with an overall lack of normal turn taking (i). The long silences (j) often are used to mark how breathtakingly inappropriate something is, for example the mathematical ignorance of a class of junior college students. The tendency to overlap (k) often takes the form of wryly sympathetic laughter. Overall, while it is often claimed that the point of a turn-taking system is to enable speakers to minimize both gaps and overlaps (Sacks, Schegloff, and Jefferson 1974; Raux and Eskenazi 2009), in this interaction style both are saliently common, with the speakers seemingly working together to do this for a specific purpose, namely to display agreement about how bad something is.

This style includes frequent use of the Positive Assessment Construction (n), but mostly used to show regard for the interlocutor’s opinion, rather than positive feeling about the subject under discussion. This style is also rich in topic continuation (l) and topic development (m), when the conversation moves to a different aspect of the same topic. In this style, these are often used when piling up evidence for an opinion, as when predicting a politician’s comeuppance with it’s just going to keep dogging him, followed by we think he’s going to be out of the race in two weeks. It is also rich in uses of the Bipartite Construction (o), the prosodic pattern often transcribable with a colon. Generally serving to adding more information to make a point clearer or stronger, in this style it was often used to stress or elaborate on the problematic nature of a situation, as in the only criterion for getting into the junior college was being 18 years of age: having a high school diploma was not a prerequisite.
While this style was generally associated with negative feelings, this was not invariable: in some fragments the speaker used it in the course of expressing grudging admiration, as in *that was pretty good . . . it was pretty clever*, but an overall positive assessment.

No topics were significantly high on Dimension 3 by our strict threshold, but there was a tendency for fragments on the topic of “AIDS” to be higher on this dimension.

Table 4
Feature Loadings of Interaction Style Dimension 3

<table>
<thead>
<tr>
<th>Feature</th>
<th>Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td>has turn</td>
<td>-.00</td>
</tr>
<tr>
<td>turn take</td>
<td>-.09</td>
</tr>
<tr>
<td>backchannel cueing</td>
<td>-.09</td>
</tr>
<tr>
<td>topic closing</td>
<td>-.09</td>
</tr>
<tr>
<td>topic development</td>
<td>.22^m</td>
</tr>
<tr>
<td>turn–hold fillers</td>
<td>.13</td>
</tr>
<tr>
<td>long turn take</td>
<td>-.07</td>
</tr>
<tr>
<td>late peaks</td>
<td>.01</td>
</tr>
<tr>
<td>meta comment</td>
<td>-.04</td>
</tr>
<tr>
<td>minor third cue</td>
<td>-.05</td>
</tr>
</tbody>
</table>

5.6 Dimension 3, Negative Pole

The negative pole, conversely, relates to a *positive assessment* stance. The prosodic feature loadings show a tendency for the participants to make use of the normal turn structuring mechanisms (*p*), including short silences (*q*). Relatively common lexical items include realistic, practical, reasonably, and frankly. More informatively, fragments near this pole generally convey positive feeling, as speakers talk for example about their dog, a great fishing day, a successful wood staining project, or their favorite football teams. Again, this style is not invariably associated with positive assessment: in one fragment the speaker used it ironically as he described an experience of being discriminated against, with his strongly negative feeling conveyed only in his word choices.

The topics significantly low on Dimension 3 were “metric system” and “space flight and exploration,” perhaps reflecting the large number of engineers among the participants, especially those volunteering to talk about these topics.

5.7 Dimension 4, Positive Pole

This pole relates to *nonfocal speaker speaks knowledgeably*, with the focal speaker taking a supportive role: acknowledging the authority or “epistemic rights” of the other (Heritage and Raymond 2005). This happens commonly when the nonfocal speaker talks at length, for example, about his woodworking hobby, his garden, school funding models, or the mechanisms of background checks.

Characteristic words include the fillers and, um, and uh, and content words such as peppers, bands, fry, gym, timing, and transmission.

Regarding the prosodic loadings, the speakers load oppositely on the turn-taking dimensions, with the nonfocal speaker slightly tending to more often hold the floor.
While the focal speaker sometimes has the turn briefly or weakly (s), they tend to readily yield it (t). Such turns are often supportive reactions (u) such as yeah and short follow-up questions of various kinds (v), often acknowledging the other as an expert on the topic, as in so why are we doing this?, which elicited a long response, starting with It's a research-type project for voice . . .

Table 5
Feature Loadings of Interaction Style Dimension 4

<table>
<thead>
<tr>
<th>Feature</th>
<th>Dimension 4, Negative Pole</th>
</tr>
</thead>
<tbody>
<tr>
<td>std deviation ranges:</td>
<td></td>
</tr>
<tr>
<td>focal has turn</td>
<td>–.03 –.05 .06* –.00 –.06 .05* .03</td>
</tr>
<tr>
<td>turn take</td>
<td>–.04 –.08 –.19* .00 .19* .08 .04</td>
</tr>
<tr>
<td>backchannel cueing</td>
<td>–.01 .01 .06 –.06 –.01 .01</td>
</tr>
<tr>
<td>empathy bid</td>
<td>.06 .13 .12 –.00 –.12 –.13 –.06</td>
</tr>
<tr>
<td>long turn take</td>
<td>.06 .10 .13 .00 –.13 –.10 –.06</td>
</tr>
<tr>
<td>late peaks</td>
<td>–.16* –.31* –.39* –.00 .39* .31* .16*</td>
</tr>
<tr>
<td>minor third cue</td>
<td>.11 .18* .26* .00 –.26 –.18 –.11</td>
</tr>
<tr>
<td>receiving action cue</td>
<td></td>
</tr>
</tbody>
</table>

5.8 Dimension 4, Negative Pole

The negative pole is, conversely, focal speaker speaks knowledgeably. Common words include nah, damn, super, hey and y'all, all of which may express dominance. In terms of the LIWC categories, fragments near this pole are high in “clout,” a category relating to leadership and confidence.

The prosodic features include a strong tendency for the focal speaker to take the turn (w). There is also a tendency to use late peaks (x) (in which a stressed syllable’s pitch peak occurs later than its energy peak), which are common in storytelling and when inviting inference.

While no topics were significantly high or low on Dimension 4, fragments on the topic “public education” tended to be further from the mean, perhaps because in many of these conversations one of the participants was an educator and thus speaking more knowledgeably.

5.9 Dimension 5

In the interest of space, we describe the next four dimensions more concisely.

The positive pole of Dimension 5 involves thoughtful style. There are no salient lexical or LIWC tendencies, but, among the prosodic feature loadings, the strongest is for the extreme Meta-Comment bin. This construction is generally used for off-topic utterances, and in these fragments the associated breathy voice often marks extreme topic shifts, appearing often on words like so and anyway, often as the conversation moves into or out of a speculative section. Long pauses are also common, as the participants consider, for example, the future of gender roles or discuss the appropriateness of a drivetrain modification.

The negative pole of Dimension 5 is a factual style. In terms of the LIWC categories, this tends to be “authentic.” Among the prosodic features, the highest loading is for the Topic Closing Construction, with the associated low pitch typically used in these fragments to express a stance of calm rationality, as the speaker describes something they know well, such as how a network is set up or how security cameras work.
There also seemed to be low tolerance for silence, with the speakers keeping going and sometimes buying time with repetitions of words or phrases.

The topics significantly high on Dimension 5 were “social change” and “capital punishment.” “Movies” and “clothing and dress” (with the prompt “find out how the other caller typically dresses for work) significantly low.

5.10 Dimension 6

For Dimension 6, the positive pole involves a envisioning positive change style. The prosodic indicators of this style are diverse, including a frequent use of the Topic Continuation Construction and a general lack of turn taking, as one speaker proceeds to tell his or her story, often involving individual choices. This pole is high on the LIWC categories of analytic, discrepancy, and work. This style is seen, for example, in talking about changes in exercise regimens, stereotypes about women’s roles, and medical research. Fragments near this pole often involve talk about individual choices, individual responsibility, and taking the initiative.

The negative pole of Dimension 6 involves a style of accepting things beyond individual control. This can involve situations like living in a small town where the big music groups never come, or a new corporate promotion policy, or the prevalence of gun-safety carelessness in the population. The prosodic tendencies are complex, but the most salient is the frequent occurrence of fairly lengthy silences. The lexical tendencies are also diverse, but common words include nope, uncomfortable, and weeds.

The topics significantly high on Dimension 6 were “care of the elderly,” and “trial by jury”, where many speakers seemed to feel that improving things could be possible. Conversely “football,” “vacation spots,” “weather climate,” “TV programs,” and “movies” were significantly low, perhaps because future developments in these areas are mostly beyond individual control.

5.11 Dimension 7

The positive pole of Dimension 7 is a style of referencing shared experiences, as the speakers seek common ground and agreement on something they know or do. This occurred with a wide variety of topics, including experiences with drug testing and with catalog shipping, and ways to make hamburger or grow snow peas, but overall the speakers were seeking something to agree about, and usually succeeding. Relatively common words include essentially, relatively, nicer, reality, deciding, and determine. The strongest prosodic tendency was to frequently use the Bookmarked Narrow Pitch Construction, which generally serves to mark contrast, for example when the speakers contrasted two James Bond actors and agreed on which they preferred, and contrasted their shared like of home-grown chillies with their respective family members’ dislike for them.

The negative pole of Dimension 7 relates to a style of making points, as the speakers stated and justified opinions, often referring to experiences, perspectives, or knowledge that the conversation partner did not share. The strongest prosodic loading was for the Bipartite Construction, which in this style is often used when listing up examples that illustrate a general point, as in if you buy through Bean or through, um, Lands End, and she don’t want to be out by herself at night anywhere in Dallas, regardless of whether it’s North Dallas, Oak Cliff, or whatnot, and information, communication and computers as harbingers of a new age. Common are words with strong valence, such as marvelous, perfectly, mentality, definite, obvious, and literally.
Table 6
Functions of the Top 8 Dimensions. The second column shows the amount of variance explained by each dimension.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Percentage</th>
<th>Description 1</th>
<th>Description 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13%</td>
<td>both participants engaged</td>
<td>lack of shared engagement</td>
</tr>
<tr>
<td>2</td>
<td>11%</td>
<td>focal speaker mostly talking</td>
<td>focal speaker listening actively</td>
</tr>
<tr>
<td>3</td>
<td>8%</td>
<td>positive assessment</td>
<td>negative feelings</td>
</tr>
<tr>
<td>4</td>
<td>5%</td>
<td>focal speaker speaks knowledgeably</td>
<td>nonfocal speaker speaks knowledgeably</td>
</tr>
<tr>
<td>5</td>
<td>5%</td>
<td>factual</td>
<td>thoughtful</td>
</tr>
<tr>
<td>6</td>
<td>4%</td>
<td>accepting things beyond individual control</td>
<td>envisioning positive change</td>
</tr>
<tr>
<td>7</td>
<td>3%</td>
<td>making points</td>
<td>referencing shared experiences</td>
</tr>
<tr>
<td>8</td>
<td>3%</td>
<td>unfussed</td>
<td>emphatic</td>
</tr>
</tbody>
</table>

The topics significantly high on Dimension 7 were “social change,” with the prompt “how is life in America different today,” which often led to talk of shared experiences, and “hobbies and crafts,” which often led the speakers to try to identify a shared interest.

5.12 Dimension 8

The positive pole of Dimension 8 is emphatic, involving a clearly conveying a strong opinion. This commonly involved either dislike, for example of fools, or people who are stupid or without common sense, or who aren’t motivated, or admiration, for example of an outstanding musical or a teaching technique where everybody might learn something. Speakers often seemed to be using this style to get a reaction from the other participant. Common words include intensifiers such as awfully and clearly, and words of strong judgment such as gorgeous and correctly. Words rare at this pole include hedges such as hopefully and generally. Prosodic characteristics include strong uses of the Enthusiastic Overlap, Topic Continuation, Positive Assessment, and Bipartite Constructions.

The negative pole of Dimension 8 is an unfussed style. Characteristic phrases include I dunno, I wonder, and who knows. Examples near this pole include talk about remote and half-understood political events, half-forgotten happenings, half-hearted ideas for political reform, and pets who died at a ripe old age. This seems to be a generally low-energy state. Prosodically this involves many constructions, notably including lengthened syllables marking an attempt to recall something or formulate an idea. Common words include routine, casual, slightly and practical; uncommon ones include excellent, properly, awfully, greatest, surprising, frankly, absolute and extremely.

On Dimension 8, the topic “child care” was significantly high, with “basketball” and “vacation spots” significantly low.

5.13 Discussion

We have thus identified meanings for 8 dimensions of style, extending our understanding of interaction style variation. Table 6 summarizes our interpretations of these top 8 dimensions. While these certainly do not constitute an exhaustive model, they do account for 53% of the variance in the features.
There is no mathematical necessity for the dimensions derived from prosodic behavior frequencies to be informative regarding interaction styles, but it turns out that they do. There is also no mathematical necessity for the interpretations of each dimension’s poles to be opposed, but for each dimension it turns out that they roughly are.

Fully independent validation of these interpretations is a topic for future work, but we did find diverse supporting evidence. First, the dimensions discovered are plausible and they explain much of the variance. In these respects, we match best practice to date in the development and validation of dimensional models of style (Biber 2004; Grothendieck, Gorin, and Borges 2011; Laskowski 2016). Second, we additionally showed that, although derived purely from prosodic features, many of the dimensions align with meaningful differences in lexical frequencies and with topic tendencies.

While we are confident in our interpretations of the dimensions, we must note that the strength of evidence varied. For Dimensions 1 and 2 there was clear evidence of all types, for Dimensions 3 and 4 there was diverse strong evidence, and for Dimensions 5 through 8 the evidence was harder to interpret, requiring more effort to find the commonalities and arrive at satisfactory descriptions.

The highest ranked dimensions are also the ones that relate best to previous research. For example, Dimension 1, low vs high engagement, aligns with Biber’s information vs interactive dimension, with Grothendieck et al.’s male vs female dimension, and with Tannen’s high-consideration vs high-contact dimension. Dimension 2, speaking vs active listening, relates to Laskowski’s time spent talking dimension. The existence of such similarities suggests that we may be approaching a general model of English interaction style variation, rather than something specific to Switchboard and this set of features.

However, no dimension here appears to exactly match any found previously. For example, information focus and low engagement seem to correlate in this corpus only weakly, and so Dimension 1 accordingly aligns only partly with Biber’s first dimension. Similarly Tannen’s third dimension, “content versus interpersonal involvement,” relates to two dimensions in the current model, engagement vs the lack thereof, and making a point vs referencing shared experiences, but again does not align perfectly with either. Here the current model may be usefully separating out independent dimensions that were previously confounded.

Our model in addition reveals dimensions that previous literature had not noted, including 5 through 8 (Table 6). Thus our method succeeded in supporting the goal of identifying more dimensions of interaction style.

6. Individual Differences

Classic user modeling work implicitly assumed that each individual has their own style, which they tend to use, and which they prefer their interaction partners to use. However, the validity of this assumption has been studied only for a handful of style aspects, and has mostly only reported on the existence of such tendencies, not how much of the variance they explain. Having now a well-defined space of interaction styles, we are able to take a broad look at this question1.

The first thing to note is that only two of the eight dimensions encode speaker-specific roles: Dimensions 2 and 4, representing who is the primary speaker and who

1 This section is largely based on (Ward 2021a).
is acting more knowledgeably. The other six are collaborative or “aligned” (Garrod and Pickering 2009), in the sense that the tendencies for both speakers are the same. For example, Dimension 3 represents the fact that generally either both speakers or neither are taking a positive stance. Even Dimensions 2 and 4 are collaborative in the sense that the less active speaker generally acts to support the main speaker in their role. This was a surprise to us, as we had originally expected to discover a space where individual differences in interaction style would be immediately evident.

We next looked at whether speakers were consistently located in the space. Among the many possible ways to quantify speaker consistency, including measuring the utility of style information for speaker identification (Laskowski 2014), we follow Weise and Levitan’s (Weise and Levitan 2020) approach, as this most directly relates to the most likely use case: an adaptive dialog system needing to predict and choose an appropriate interaction style for an upcoming dialog. The use of speaker information should enable more accurate predictions, so we measure the extent to which this is true.

Our vector space representation enables us to measure the distance between any two interaction styles, and in particular, between a predicted style and the observed style. Distance is how we evaluate prediction quality: we use the mean squared difference for each dimension, and also the Euclidean distance across dimensions.

Of course this is only valid if fragments closer in this space are also perceived as closer in style. A proper investigation of perceptions is beyond the scope of this paper, but we did spot check a few of the pairs that were closest in the reduced, 8-dimensional space, and found that each pair was indeed very similar in style. Moreover, these pairs appeared across diverse regions of the space, and the proximity-similarity connection held in every case; for example, for reminiscing about childhood situations that were annoying at the time but now seem nostalgic, with the interlocutor supportively showing empathy based on similar experiences; for jumping right in to address the assigned topic with a near monologue, with the interlocutor just occasionally chiming in with agreement; and for explaining political or corporate policies that the interlocutor is also familiar with and views in the same way. (Incidentally, we also observed the same connection when computing the distance using all 84 dimensions.)

As a baseline model of style predictability, we just predict the global average style for every fragment. The model exploiting individual information predicts the interaction style as the average of the interaction styles in other fragments with one of the participants, excluding fragments from the same dialog. The models were evaluated using only fragments for which the 33022-fragment subset included at least 20 others by the same speaker in different conversations, that is, at least 10 minutes of reference data to use for independent estimation of the individual’s style. There were 31931 such fragments.

<table>
<thead>
<tr>
<th>predictor</th>
<th>dimension 1</th>
<th>dimension 2</th>
<th>dimension 3</th>
<th>dimension 4</th>
<th>dimension 5</th>
<th>dimension 6</th>
<th>dimension 7</th>
<th>dimension 8</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>speaker’s average style</td>
<td>5.8%</td>
<td>4.0%</td>
<td>17.0%</td>
<td>2.5%</td>
<td>5.3%</td>
<td>8.0%</td>
<td>0.5%</td>
<td>2.7%</td>
<td>3.57%</td>
</tr>
<tr>
<td>gender average style</td>
<td>0.6%</td>
<td>0.0%</td>
<td>0.6%</td>
<td>0.0%</td>
<td>1.2%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.4%</td>
<td>0.21%</td>
</tr>
<tr>
<td>age-range average style</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.3%</td>
<td>0.3%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.4%</td>
<td>0.0%</td>
<td>0.06%</td>
</tr>
</tbody>
</table>

Table 7
Average prediction error reductions for various models: reductions per dimension in mean squared error and reductions overall in Euclidean distance, all relative to the baseline.
The first row of Table 7 shows the reductions in prediction error obtained using the individual models, compared to the global-average baseline. Overall, knowing the speaker’s identity reduces the average prediction error by only 3.6%, a rather modest amount. (The benefit was even less if the distances are computed using all 84 dimensions: only a 1.8% reduction.)

This was a surprise to us; it implies that the styles are not very stable: that individuals vary greatly in style. This is, however, not hard to understand; in real life, we know that how people talk varies with the situation, topic, interlocutor, time of day, and other factors. Still, to put this in perspective, it implies that, even though the participants in this corpus are generally good conversationalists, a system able to miraculously adapt as well to the interlocutor’s average style would perform only 3.6% better than one which did not.

7. Other Findings

This section reports a few small follow-on explorations.

7.1 Variation in Predictability

While individual style consistency was weak, for some dimensions it was stronger. The per-dimension reductions in prediction error (Table 7) are largest for Dimensions 3 and 6, suggesting that individuals tend to be relatively consistent for the negative vs positive dimension and for the accepting vs progress-oriented dimension.

We investigated also whether predictibility would be easier for later fragments. Since entrainment in general takes time (Wynn and Borrie 2020), we thought that fragments taken from later into the calls might be closer to each participant’s “true” style, as he or she came to reveal and relax into their own preferred style, and discover and compromise towards their partner’s preferred style. We therefore hypothesized that the styles of later fragments would be more predictable, but this turned out not to be the case: rather there was a slight tendency to less typical behavior over time.

While knowledge of individual styles was not very informative in general, this varied across speakers. Speaker knowledge enabled better predictions for 78% of the speakers, while for the remainder, the global average was more predictive. For the most predictable speaker, the mean distance for predictions was only 50% of the average (she consistently took a passive listening role). Others were highly unpredictable, including one whose distance was over 4 times the average. Listening to some of her calls, to infer what factors might cause greater variability, we noticed that she was calling at different times during the day, sometimes had a baby crying in the background, and participated in conversations about a wide variety of topics.

7.2 Demographic Differences

Interaction style differences among populations are a popular topic of discussion, but previous work has often relied on anecdotal evidence. With this model in hand, we are equipped to study differences quantitatively. To the extent that subpopulations have characteristic interaction styles, we can expect that knowing the group identity of a speaker would be informative regarding the interaction style in fragments including them. The second and third rows of Table 7 show the results when predicting using two other types of knowledge: the speaker’s gender and their age range, above or below 38 years old, the mean for this corpus. Predictions based on gender are only about 0.2%
better than generic predictions, and the age-class predictions show even less benefit. Clearly the variation within these subpopulations is hugely greater than the variation between them.

Since women are often said to take more of the burden of adapting to their interlocutor, we hypothesized that they would tend to vary more in style. The average prediction error reduction obtained by using the individual models for women was 2.1% and for men 6.1%, so the women did indeed diverge more from their average styles.

Although the subpopulation tendencies had little predictive power, it is interesting to consider what they suggest. Table 8 shows the means for four splits of the 33022 fragments: by gender, by age group, by order of joining the call, and by time into the call. Statistically, fragments with women participating tend to be more engaged, negative, and factual in style (Dimensions 1, 3, and 5, effect sizes .16, .16, and .22 standard deviations, respectively). Fragments with the older speakers also tend to be more negative in style, and the older speakers tend to a more knowledgeable style (Dimensions 3 and 4, .13 and .10). Fragments later in the conversation, specifically those occurring after 4 minutes in, also tend to be more negative in style (.14). The speaker who joined the conversation first tended slightly to talk more and to act more knowledgeably (Dimensions 2 and 4, .04 and .05), which makes sense, as they were instructed by the robot operator to “Please think about the topic while I locate another caller” (Godfrey, Holliman, and McDaniel 1992), which sometimes took a few minutes. All of these differences are statistically significant ($p < 0.0005$, two-sided, unmatched-pairs, t-tests, with Bonferroni correction), but the effect sizes are small.

These findings are only suggestive. For one thing, the Switchboard corpus was not collected to support population comparisons. For another, they reflect only prosodic behaviors, so, for example, the apparent tendency to negativity for women could be balanced out by a tendency for women to be lexically more positive. Nevertheless, they illustrate how this model can be used to study population differences.

Table 8
Average Dimension Values for Various Data Subsets

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>0.26</td>
<td>-0.06</td>
<td>-0.20</td>
<td>-0.02</td>
<td>0.22</td>
<td>0.05</td>
<td>0.04</td>
<td>-0.09</td>
</tr>
<tr>
<td>female</td>
<td>-0.28</td>
<td>0.06</td>
<td>0.21</td>
<td>0.02</td>
<td>-0.23</td>
<td>-0.06</td>
<td>-0.04</td>
<td>0.09</td>
</tr>
<tr>
<td>young</td>
<td>-0.06</td>
<td>0.08</td>
<td>-0.13</td>
<td>0.09</td>
<td>-0.00</td>
<td>-0.03</td>
<td>0.08</td>
<td>-0.02</td>
</tr>
<tr>
<td>old</td>
<td>0.09</td>
<td>-0.11</td>
<td>0.18</td>
<td>-0.12</td>
<td>0.00</td>
<td>0.05</td>
<td>-0.12</td>
<td>0.03</td>
</tr>
<tr>
<td>earlier</td>
<td>0.05</td>
<td>0.00</td>
<td>-0.14</td>
<td>-0.00</td>
<td>0.08</td>
<td>0.01</td>
<td>0.05</td>
<td>-0.06</td>
</tr>
<tr>
<td>later</td>
<td>-0.08</td>
<td>-0.00</td>
<td>0.21</td>
<td>0.00</td>
<td>-0.13</td>
<td>-0.02</td>
<td>-0.07</td>
<td>0.09</td>
</tr>
<tr>
<td>first</td>
<td>0.00</td>
<td>-0.07</td>
<td>0.00</td>
<td>-0.06</td>
<td>-0.00</td>
<td>-0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>second</td>
<td>-0.00</td>
<td>0.07</td>
<td>-0.00</td>
<td>0.06</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.00</td>
</tr>
</tbody>
</table>

8. Summary, Implications, and Future Directions

The first contribution of this work is a workflow for discovering, from data, the dimensions of variation in interaction styles. Our key innovation is in the feature set, which captures many aspects of dialog behavior. This workflow enables automatic comparison
of the styles of specific conversations and will support future work on dialog systems adaptation.

The second contribution of this work is a description of 8 important dimensions of style variation in American English, as summarized in Table 6. Thanks to the use of more features and more data, this extends our understanding beyond the handful of style aspects identified by previous work. This opens the way to richer description of conversations and corpora, and can be used in support of data analysis, corpus selection, system design and adaptation, and the other the purposes discussed in Section 2, used either qualitatively or quantitatively.

The third contribution of this work is the finding that individual styles explain little. This was seen in the fact that most aspects of style were engaged in collaboratively, with the participants taking on different roles in only 2 of the 8 dimensions, and in the fact that individual style tendencies accounted for only 3.6% of the variation. This suggests that future work on style adaptation for spoken dialog systems should not prioritize adaptation to the user, but rather adaptation to other factors such as the situation, topic, and dialog activity type.

Other surprises include the finding that gender explains very little of the variation in interaction styles, and the finding that the most stable aspect of speakers is their tendency to a positive vs. negative style.

There are many questions for future work to explore.

Further work should examine the perceptual space of interaction styles. One might, for example, measure how well similarity in this space matches human similarity judgments, or explore how people perceive these dimensions of style. While we think that most people are adept at subconsciously detecting and adopting various interaction styles, we know that elucidating this will require clever experiment design, since in several small pilot studies we found people to be poor at explicitly judging or describing interaction style dimensions. Individual variation in perceptions will also be an issue (Ranganath, Jurafsky, and McFarland 2013).

Future work might examine robustness with respect to different feature sets and analysis methods. In several small-scale experiments, some inadvertent, we built models in different ways — using features with different bin ranges, fewer features, more features including some word-class features, 32-second fragments as datapoints, and full conversations as datapoints — and, while the resulting dimensions were never exactly the same, they were always similar. Further work might examine this systematically.

Future work might also examine the generality of this model across data sets. Mathematically, the 8 dimensions found here are one of infinitely many partial spanning sets for the space of interaction styles. For any specific purpose, a different set will likely be better, in the sense of explaining more variation or doing so with fewer dimensions, and the optimal set will doubtless vary from one data set to another. Nevertheless, given Biber’s work suggesting that any reasonably diverse corpus can serve as a useful microcosm for all the genres of a language, our 8 dimensions may have general utility. At the same time, we do not believe that there are only 8 important dimensions. Here we stopped after 8 because we had no success interpreting the lower dimensions. While more sophisticated modeling techniques might help, we feel the most promising direction for extending our understanding would be to apply these methods to larger, multi-genre data collections.

Future work should also examine individual preferences. Although people in these conversations exhibited a variety of styles, it is possible that, as users, they would prefer a dialog systems to consistently use a fixed, individually-congenial interaction
style. Examining this might further lead to individualized models of the mapping from system behavior to satisfaction properties (Yang, Levow, and Meng 2012).

Despite the many open questions, the model as-is may have value. For example, the new dimensions found can suggest additional style control parameters that speech synthesizers and text generators will in future need. Further ahead, this model and these techniques may support the faster creation of better dialog systems, more focused computational pragmatics studies, and better communicative-effectiveness coaching for individuals.

To support further investigations and applications, the code for our workflow and models is available at https://github.com/nigelgward/istyles.

9. Acknowledgments

We thank Aaron Alarcon for feature extraction code for a preliminary investigation, and David Novick, Olac Fuentes, and Yuko Ward for discussion.

References


Ward & Avila

Interaction Styles


Niebuhr, Oliver. 2014. Resistance is futile: The intonation between continuation rise and calling contour in German. In Interspeech, pages 132–136.


Raux, Antoine and Maxine Eskenazi. 2009. A finite-state turn-taking model for spoken
dialog systems. In *NAACL HLT*.


