A Dimensional Model of Interaction Style Variation in Spoken Dialog

Nigel G. Ward, Jonathan E. Avila

Department of Computer Science
University of Texas at El Paso
500 West University Avenue
El Paso, TX 79968-0518
nigelward@acm.org

March 8, 2023

In spoken dialog people vary their interaction styles, and dialog systems should be able to do the same. Previous work has elucidated various aspects of style variation and adaptation, but a general model has been lacking. Here we present a dimensional model of the space of interaction styles, derived from a large data set and prosody-based features. The 8 dimensions of this model cover many previously-noted aspects of style and include some novel ones. This model may be useful for selecting data for dialog model pretraining and fine-tuning, for investigating demographic differences, and for dialog system style adaptation. However, regarding individual differences in interaction style, we find individual style tendencies to be surprisingly weak, with a predictive model based on individual tendencies outperforming a speaker-independent model by only 3.6%.

Keywords: conversational styles, dialog activities, prosody, turn taking, vector space model, individual differences, corpus characterization

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, fast-paced, playful, chatty, soft-spoken, businesslike, domineering, and many more. Styles vary among groups among individuals, across dialogs, and even within dialogs, as the topics, mood, and speakers' goals change. While previous research has diversely identified many aspects of style, for both practical and scientific reasons we need a consistent general model of the space of interaction style variation.

This paper describes such a model, created by applying Principal Component Analysis over a large set of features that represent behaviors in spoken dialog. Section 2 explains why we need a good model of interaction style. Section 3 discusses previous
linguistic, psychological, and computational approaches to interaction styles. Section 4 describes how we used Principal Component Analysis over the Switchboard corpus, using a 84 novel features that encode the frequencies of diverse interaction-related prosodic behaviors, to derive a model and to interpret its dimensions. Section 5 describes the manifestations and meanings of the top 8 dimensions of interaction style. Section 6 illustrates how the model can be used to investigate subcorpus differences, gender and age differences, and individual differences, and reports the surprising result that individual interaction style tendencies were mostly very weak. Finally Section 7 summarizes and discusses prospects.

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. After inferring a user’s interaction style, and their preferences regarding style, a dialog system should be able to adjust its behavior to suit, in order to improve efficiency and acceptability for each specific user [1, 2]. To date, most dialog systems have an unvarying, bland style. This can be attributed to the aim of working acceptably for almost any user, rather than very well for any specific user, and to the goal of covering for technology limitations, typically by adopting a fairly formal, dominant, and slow-paced style, to guide user expectations and behaviors. However today the core technologies have advanced to the point where style variation is becoming realistically possible. One exemplar is Metcalf’s adaptation of Siri to detect which users are more talkative and then provide them information in a more chatty style [3]. Building this system required a specific insight and much careful engineering, but ultimately such adaptations should be commonplace and straightforward. To support this we need a broader understanding and a general model of the space of interaction styles.

Interaction style control is also an issue for multifunctional dialog systems, such as those capable of both task-oriented and chitchat interaction [4], or more diverse variation, where, to avoid incongruity, the style may need to change to suit the current dialog activity [5].

Accordingly our research questions include: 1) What are the main aspects of style, beyond a few obvious ones such as talkative/quiet and businesslike/casual? 2) What aspects of style variation are most common in everyday interaction? 3) How can we integrate previous descriptions of style aspects into a general model? 4) How stable are individuals’ interaction styles?

2.2 Systematizing Corpus Description

Ideally, the development of dialog systems would be an engineering discipline based on a solid scientific understanding of the pragmatics of dialog. Today, however, that foun-
Dation is weak. One reason is that the field of pragmatics has been only sporadically cumulative: it is dismayingly common for reports of a correlation between some measure and some percept 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. Much of the confusion is likely due to unremarked differences in the styles of the corpora used in these studies. 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 [6]. But many 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 [7]. A better understanding of corpus differences could support the resolution of such apparent contradictions [8]. Today, alas, this is not possible, and indeed, most published descriptions of corpora specify only 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 resolve apparent contradictions, and ultimately map out the field of inquiry of computational pragmatics, enabling us to advance from a collection of scattered observations across unrelated corpora to a unified understanding.

Poorly understood corpus differences impede not only pragmatics as a science, but also the collection and selection of corpora to support dialog systems development. Today dialog systems developers rely heavily on machine learning from corpora to acquire the various knowledge and parameters needed for an effective system, but the process of selecting corpora is rarely systematic. Existing methods for corpus comparison rely mostly on lexical and topic overlap, so it is difficult to predict how well other types of knowledge will transfer. As a result, it is not uncommon for a new development project to start with the collection of a domain-specific corpus. Instead, we would like to be able to better leverage existing corpora [9]. One recent success was a recommendation system for movies that was created by exploiting the subset of Switchboard data relating to movies [10]. Selecting this subset was easy 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, even when annotations are lacking.

The need here is increasing with the trend towards large corpora designed to broadly support dialog systems design or the creation of pretrained models, where we wish to be able to make informed choices of the corpora on which to pretrain. Today this is largely based on convenience and hunches. For example, several recent large efforts have aimed to collect dialogs to exemplify a sweet-spot style that is simultaneously natural for humans and implementable with current technology [11, 12]. These efforts have used clever manipulations such as having dialogs be written down and then read, or interposing a text-to-speech engine between the confederate and the subjects. But we currently lack any ability to model how such manipulations affect style, let alone quantify in what respects such datasets match the needs of systems designers and developers. A reference model of interaction styles would help model how such manipulations affect style, how well the results match various application needs, and what supplementary datasets might be needed. For example, we imagine a day when models of interaction style will
support 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.

Thus our final research question: 5) How can we automatically characterize the style of any corpus?

3 Previous Work

Style has long been a topic of interest in linguistics, with a very diverse set of style properties identified and studied [13]. Most work has focused on sentence- or utterance-level styles alone, and most computational work on style has focused on style and style transfer in text generation or speech synthesis alone. Today these aspects of style are seeing rapid advances generation [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24], but their application appears to have lagged. This may be because each addresses only some components of the larger problem: that of attaining overall interaction styles that work for the task, domain and user. It has also been noted that one-by-one adaptation of individual elements of interaction style, without considering interactions among elements, risks creating dispreferred or inauthentic system behavior [22]. Accordingly, this paper focuses on these overall styles and how they are realized. We hope that our general model will help guide the development of component technologies in directions useful for applications.

Style has received rather less attention at wider timescales and in relation to phenomena of dialog and interaction, but there is still significant groundwork for us to build on. As a terminological aside, work in this area has used various terms, including “conversational styles,” “interpersonal styles,” and “conversational text types,” reflecting somewhat different areas of emphasis, but much of it is very relevant to our aims and research questions. We will continue to use the term “interaction styles,” to emphasize the centrality of focus on the properties of the interactions per se.

One research tradition in this area focuses on intergroup differences. In particular, Tannen and others have noted that these can be a major cause of social awkwardness: 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 [25, 26, 27, 28]. This is an active and complex area of research, not least because aspects of style also relate to topic [29] and stance taking [30, 31, 32], as well as to factors active over time scales far longer than one conversation, such as genre [33, 34], personality [35, 36, 37, 38], culture [39], social role, self presentation, and social identity [40].

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 [41], sometimes filtered through the lens of system or user personality [42, 43, 44], and are increasingly relevant in the customization and adaptation of spoken dialog systems [45, 46, 47, 48, 3]. These
dimensions are useful and sometimes quantifiable [49], but unfortunately do not provide
a systematic or general model of the space of interaction styles.

Another research tradition focuses on individual differences. It is evident that people
in dialog are flexible, and often adapt their behavior to their conversation partner [50].
People, and 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 individ-
ual’s speech, such as counts of different word classes, average pitch, and average articu-
lar precision [51, 52, 53, 54, 55, 56, 57]. Unfortunately, these findings do not reliably
generalize across corpora [58], and again do not provide a systematic model.

Thus there is, on the one hand, a mostly qualitative tradition that looks at “macro” as-
pects of style, aligning with socially meaningful practices, each often involving multiple
indicators, and, on the other hand, a quantitative tradition that looks at directly measur-
able, “micro” features of dialog. Of course, there are commonalities in the phenomena
studied by the two traditions: for example, overlapping speech can be treated as either a
low-level feature or a macro-level style parameter.

A third research tradition connects the two perspectives, using the correlation pat-
terns among micro features to discover macro dimensions of variation. Deferring discus-
sion of methods until the next section, here we note that work in this tradition has pro-
duced systematically-derived lists of interaction-style dimensions: Biber [33] found three
dimensions: information-focused vs interactive discourse, stance-focused vs context-
focused discourse, and narrative-focused vs present-tense discourse. Grothendieck et
al. [59] 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
[60]. Each of these is indeed a systematic model of style variation. Unfortunately none
is entirely adequate, for several reasons. None identified more than 3 dimensions, which
seems a priori inadequate for describing the wide range of observed variation (indeed,
accounting for only 36% of the observed variance explained, for the one model for which
this was reported [33]). None resulted in the release of software or models. Further,
perhaps for these reasons, none of these models appears to have found practical use in
dialog systems development.

Thus previous work provides useful observations, features, and methods, and a long,
diverse list of interaction styles and style properties, as suggested by Table 1. This paper
builds on these to further explore the space of interaction styles, including by identify-
ing and quantifying more dimensions of interaction style and by addressing our other
research questions.

4 Methods

Accordingly we aim to produce a general, systematic model of variation in interaction
styles.

To build this model, our strategy follows that of Biber [33], in that we apply Principal
Table 1: Some interaction styles and style properties. Properties that are alternatives or serve as pairs to define a dimension are shown on one line with a slash. (The citations given are representative, not necessarily the earliest, clearest or most influential discussion of each property. There are many other terms in the literature which appear to refer to a styles that is similar or very similar to one on this list; these are not included. We also omit terms that, while somewhat relevant to interaction styles, are more directly related to personality, (text) writing styles, (monolog) speaking styles, emotion, social roles, or measurements of individual style features.)
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. [60, 59], in that 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 [69], we here follow most work in this area, dating back to the earliest work by Tannen [68], 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 [33]. However, at least for spoken conversation, such features have so far revealed only three clear dimensions, which seems quite inadequate in light of the variety of styles noted by qualitative research (Table 1). (However transcript-based features may be more informative for other languages, such as Japanese, which has a wide variety of inflections and particles that convey social meanings [5].)

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 [38, 59, 60]. 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
4.2.1 Feature Design Strategy

Having chosen to use features based on prosody, there are still many choices. In theory it might be 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 — but to avoid a combinatorial explosion, we instead choose 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 [70], 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. Finding such configurations can be a challenge in itself [71], but, fortunately, for English prosody there is already a convenient inventory of the most common ones. The next subsection explains.

![Figure 1: Approximate temporal domains of influence of the components of an instance of the Positive Assessment Construction, with times in milliseconds.](image)

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 [72, 73, 74].

A simple example is the Positive Assessment Construction, illustrated in Figure 1 [75]. 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 syn-
Backchanneling Construction

<table>
<thead>
<tr>
<th>Functions:</th>
<th>encourage continued listening; encourage continued talk, etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form:</td>
<td>timespan prosodic properties, etc.</td>
</tr>
<tr>
<td></td>
<td>Speaker 1 Speaker 2</td>
</tr>
<tr>
<td>–600 ms</td>
<td>loudness, creakiness, and articulatory precision increase</td>
</tr>
<tr>
<td>–400 ms</td>
<td>pitch drops and stays low</td>
</tr>
<tr>
<td>–300 ms</td>
<td>loudness decreases</td>
</tr>
<tr>
<td>–100 ms</td>
<td>silence</td>
</tr>
<tr>
<td>0ms</td>
<td>backchannel: lengthened, quiet, flat pitch</td>
</tr>
<tr>
<td>800 ms</td>
<td>resumes speaking, loud, fast, high pitch</td>
</tr>
</tbody>
</table>

Figure 2: The Backchanneling Construction.

tactic 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 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 not as a mere superficial measure of whether two voices are simultaneously active, which can happen in many ways, 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.
1 focal has turn … other has turn Turn Hold
2 silence … enthusiastic overlap
3 turn take by focal … turn take by other Turn Exchange
4 other backchanneling … focal backchanneling Backchanneling
5 topic closing … topic continuation
6 topic development … positive assessment
7 empathy bid by focal … empathy bid other Empathy Bid
8 turn-hold fillers … bipartite construction
9 long turn take … long turn yield Long Turn Exchange
10 late peaks … turn-initial fillers
11 meta comment … bookended narrow pitch
12 minor third cue … response to an action cue Minor Third

Table 2: The Prosodic Constructions Used in the Feature Computations

At this point we need to introduce some terminology: “focal participant” and “non-focal participant.” Many interactive behaviors are symmetric, in the sense that the participants play complementary roles. For example, if one participant is holding the turn, the other participant is allowing the turn. To make it clear who’s who, we call one speaker the focal participant, so we can then, for example, describe a dialog fragment as characterized mostly by the “focal participant has turn.” By doing this, we can later indicate whether, for example, in certain situations the focal speaker is also frequently making bids for empathy, or whether it’s the nonfocal participant (the other participant) who tends to make bids for empathy. Of course “focal participant” is not a fixed role, just one that we ascribe temporarily to a participant in order to comprehensively describe which behaviors are theirs.

The constructions used in this study are listed in Table 2, using short names that evoke the most common functions of each configuration. As described in [74], they were derived automatically, from data. As various previous work indicates that many important interactive behaviors and prosodic behaviors happen within 3-second timeslices, our derivation process started by computing 106 speaker-normalized prosodic measures per speaker, for a total of 212, spanning 3.2-second windows, to broadly characterize the prosodic and turn-taking situation within each such window. This was done repeatedly, for hundreds of thousands of windows, centered at 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 joint behaviors [74]. The result was several significant dimensions of variation. Each dimension has a positive pole and a negative pole, and each of these was interpreted as a meaningful prosodic construction. Some construction pairs are “degenerate;” for example, the “turn take by focal participant” and “turn take by other participant” constructions are both instances of the Turn Exchange Construction, differing only the roles of the two speakers. Such degenerate cases are indicated in Table 2 by the presence of a general construction name in the last column. Most of these involve joint behaviors by the two participants.
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 [76, 77, 75]. 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 [78] — and used for a context-dependent set of related purposes. Indeed, the names are post hoc and not causal in any sense, as the learning of these configurations was entirely unsupervised.

These constructions were generated from one specific corpus with one specific set of measures [74], but experiments with other corpora and other measures yield fairly similar inventories [79, 80]. 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 [81]. Finally, they cover a wide range of dialog states, activities, and events, including many of those often considered most important in human interaction [82, 83], and they involve most of the low-level features examined in previous 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 these is no evidence for it indicates the prevalence of lack of engagement. Thus, rather than averaging values, which would blur away the details, we need statistics on the distribution of values, over each fragment, for each construction.

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 used to create a kind of histogram, consisting of the counts 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 2.

Figure 3 overviews this process. This method effectively pools 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. While this process is susceptible to inaccuracies due to noise or non-modeled aspects of prosodic behavior, some degree of robustness is provided by the use of 84 somewhat-redundant features: any single errorful value is unlikely to drastically affect the result for any fragment.

4.3 Data and Processing

For analysis we chose the Switchboard corpus of American English telephone conversations [84, 85], 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 [86]. As interaction styles are neither instantaneous nor stable over long times, we needed to choose a fragment size long enough to support meaningful feature computations, yet short enough that many fragments would exhibit only one style. We chose to use 30-second fragments as the unit of analysis, which gave us approximately \((30\text{ s} - 3.2\text{ s}) / .02\text{ s} = 1340\) samples per fragment; this seemed to work well. 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 participant” and B as the “nonfocal participant,” then again with the roles reversed. This enabled us to analyze both participants from both perspectives, enabling us later to discuss connections to gender and other differences. This yielded \(2 \times 16511 = 33022\) data points. This is an order of magnitude larger than in any previous attempt to model interaction styles [33, 59, 60].

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 [86]. Addressing Research Question 5, this provides a straightforward way to characterize new data using the existing BFFs, as seen at the bottom of Figure 4. Alternatively, for example for analysis of languages other than English, this can be used to 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. This representation may be useful as-is for many purposes, such as computing proximity as a proxy for similarity. However, to address Research Question 1, we went on to use these dimensions to discover meaningful aspects of style.
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/](http://www.cs.utep.edu/nigel/istyles/).
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 some 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 10th percentile on the dimension, and for the positive pole, fragments above the 90th 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 scanned the frequency ratios for the 40 most frequent words. We also examined word-class tendencies using LIWC [87], a well-known program that enables approximate inference of psychological states from word use frequencies.

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

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

Interpreting a pole was not always easy: for some it took significant effort to arrive at a satisfactory interpretation. For some poles, some types of evidence were scant or difficult to interpret, so we had to rely more on the other types. Another problem was non-conforming exemplars. Although examination of the exemplar fragments was reliably informative, for most poles there were one or more exemplars 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, but some simply seemed not to fit the general pattern. Another problem was the huge num-
Table 3: 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>Function Description</th>
<th>Interpretation</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 participant mostly talking ...</td>
<td>focal participant 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 participant speaks knowledgeably ...</td>
<td>nonfocal participant 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>

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 [86].

5 Eight Dimensions of Interaction Style

This section discusses both poles of each of the 8 dimensions of interaction style, for a total of 16 style aspects. Table 3 provides a preview.

While the first two dimensions are clearly not original, the others indicate that there do indeed exist aspects of style beyond those identified in previous work, providing an answer to Research Question 1. The amounts of variance explained suggest an answer to Research Question 2: on the assumption that the dimensions that explain more of the observed variance are the ones that are more significant in human communication, the dimensions are listed in order of importance.
This rest of this section goes into detail: describing the surface manifestations of each pole of each dimension, explaining the inferred meanings, and summarizing the evidence for the correctness of those meanings.

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 4. 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>
<thead>
<tr>
<th>Feature</th>
<th>Dimension 1 Positive Pole Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td>focal has turn</td>
<td>(-0.02)</td>
</tr>
<tr>
<td>silence</td>
<td>(0.09)</td>
</tr>
<tr>
<td>turn take</td>
<td>(-0.13)</td>
</tr>
<tr>
<td>backchannel cueing</td>
<td>(-0.12)</td>
</tr>
<tr>
<td>topic closing</td>
<td>(-0.02)</td>
</tr>
<tr>
<td>topic development</td>
<td>(-0.05)</td>
</tr>
<tr>
<td>empathy bid</td>
<td>(-0.05)</td>
</tr>
<tr>
<td>turn–hold fillers</td>
<td>(0.02)</td>
</tr>
<tr>
<td>long turn take</td>
<td>(-0.12)</td>
</tr>
<tr>
<td>late peaks</td>
<td>(-0.11)</td>
</tr>
<tr>
<td>meta comment</td>
<td>(0.08)</td>
</tr>
<tr>
<td>minor third cue</td>
<td>(-0.12)</td>
</tr>
</tbody>
</table>

Table 4: 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 4, 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 fragments 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 clarify which behaviors align with which role, we’ll refer to one participant as the “focal” participant, and the other as “nonfocal,” as before. Once again, this is not a fixed role, 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 participant listening actively, and the nonfocal participant mostly talking. The evidence is, again, diverse and consistent.

The prosodic loadings, seen in Table 5, indicate that here the nonfocal participant mostly has the turn (e). The speakers frequently perform a weak turn exchange, with the nonfocal participant yielding to the focal participant (f). The focal participant also tends to be prolifically backchanneling (g). Evidently at this pole the focal participant is in a supporting role, with the nonfocal participant doing most of the talking. The nonfocal participant 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 participant, 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 participant — for example, regarding local manufacturing, German perceptions of America, or a local criminal case — with the focal participant actively listening. In particular, the focal participant’s weak turn taking actions (f) are often briefly supportive statements, such as sure, exactly, and oh, I felt that way too.
std deviation ranges: –2.4 –1.4 –0.4 0.4 1.4 2.4
focal has turn –.01 –.17 –.28 −.00 .28e .17e .01 other has turn
turn take –.02 .01 .20f .00 −.20 −.01 .02 turn yield
backchannel cueing −.08 −.18 −.27 .00 .27g .18g .08 backchannelling
empathy bid −.06 −.17 −.28 .00 .28h .17h .06 indifference
long turn take .03 .10 .21 .00 −.21 −.10 −.03 long turn yield
late peaks −.02 −.02 .03 .00 −.03 .02 .02 turn–initial fillers
minor third cue .07 .15 .19 .00 −.19 −.15 −.07 receiving action cue

Table 5: 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 participant mostly talking, with the nonfocal participant being an active listener. Words characteristic of the focal participant 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 6) 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, despite the
The general claim is that the point of a turn-taking system is to enable speakers to minimize both gaps and overlaps \cite{88, 89}. In this interaction style, both patterns are saliently common, with the speakers seemingly doing this together 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 in \textit{it's just going to keep dogging him}, followed by \textit{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 add 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 \textit{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 invariably: in some fragments the speaker used it in the course of expressing grudging admiration, as in \textit{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>
<thead>
<tr>
<th>Feature</th>
<th>Dimension 3 Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td>std deviation ranges:</td>
<td>–2.4  –1.4  –0.4  0.4  1.4  2.4</td>
</tr>
<tr>
<td>has turn</td>
<td>–.00  –.14(p)  .04  .12(i)  .04  –.14(p)  –.00</td>
</tr>
<tr>
<td>silence</td>
<td>.15(i)  .02  –.26(q)  .10  .17(k)  .16(k)  .11(k)</td>
</tr>
<tr>
<td>turn take</td>
<td>–.09  –.08  –.04  .12(i)  –.04  –.08  –.09</td>
</tr>
<tr>
<td>backchannel cueing</td>
<td>–.09  –.10(p)  –.01  .13(i)  –.01  –.10(p)  –.09</td>
</tr>
<tr>
<td>topic closing</td>
<td>.04  –.02  –.09  –.02  .06  .12(f)  .13(f)</td>
</tr>
<tr>
<td>topic development</td>
<td>.22(m)  .26(n)  –.12  –.25  .17(n)  .27(n)  .18(n)</td>
</tr>
<tr>
<td>empathy bid</td>
<td>–.00  –.03  –.01  .07  –.01  –.03  –.00</td>
</tr>
<tr>
<td>turn–hold fillers</td>
<td>.13  .20  –.06  –.23  .12  .22(o)  .15(o)</td>
</tr>
<tr>
<td>long turn take</td>
<td>–.07  –.10  –.03  .12(i)  –.03  –.10  –.07</td>
</tr>
<tr>
<td>late peaks</td>
<td>.01  –.01  –.03  .07  –.03  –.01  .01</td>
</tr>
<tr>
<td>meta comment</td>
<td>–.04  .00  .09  .04  .02  .01  .01</td>
</tr>
<tr>
<td>minor third cue</td>
<td>–.05  –.08  –.03  .14(i)  –.03  –.08  –.05</td>
</tr>
</tbody>
</table>

Table 6: Feature Loadings of Interaction Style Dimension 3

### 5.6 Dimension 3, Negative Pole

The negative pole, conversely, relates to a \textit{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 \textit{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 participant speaks knowledgeably, with the focal participant taking a supportive role: acknowledging the authority or “epistemic rights” of the other [90]. This happens commonly when the nonfocal participant 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 participant slightly tending to more often hold the floor (r). While the focal participant 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>
<thead>
<tr>
<th>std deviation ranges:</th>
<th>–2.4</th>
<th>–1.4</th>
<th>–0.4</th>
<th>0.4</th>
<th>1.4</th>
<th>2.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>focal has turn</td>
<td>–.03</td>
<td>–.05</td>
<td>.06*</td>
<td>–.00</td>
<td>–.06</td>
<td>.05*</td>
</tr>
<tr>
<td>turn take</td>
<td>–.04</td>
<td>–.08</td>
<td>–.19*</td>
<td>.00</td>
<td>.19*</td>
<td>.08</td>
</tr>
<tr>
<td>backchannel cueing</td>
<td>–.01</td>
<td>.01</td>
<td>.06</td>
<td>–.00</td>
<td>–.06</td>
<td>–.01</td>
</tr>
<tr>
<td>empathy bid</td>
<td>.06</td>
<td>.13</td>
<td>.12</td>
<td>–.00</td>
<td>–.12</td>
<td>–.13</td>
</tr>
<tr>
<td>long turn take</td>
<td>.06</td>
<td>.10</td>
<td>.13</td>
<td>.00</td>
<td>–.13</td>
<td>–.10</td>
</tr>
<tr>
<td>late peaks</td>
<td>–.16*</td>
<td>–.31*</td>
<td>–.39*</td>
<td>–.00</td>
<td>.39*</td>
<td>.31*</td>
</tr>
<tr>
<td>minor third cue</td>
<td>.11</td>
<td>.18*</td>
<td>.26*</td>
<td>.00</td>
<td>–.26</td>
<td>–.18</td>
</tr>
</tbody>
</table>

Table 7: Feature Loadings of Interaction Style Dimension 4

5.8 Dimension 4, Negative Pole

The negative pole is, conversely, focal participant 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 participant 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. Con-
versely “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 shopping, 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 chilies 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.

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

Research Question 1 asked about the main aspects of style: these 8 dimensions provide a new answer, as summarized above in Table 3. While these certainly do not constitute an exhaustive model, they do account for 53% of the variance in the features and significantly extend our understanding of the space of interaction styles.

While we leave proper evaluation of these interpretations for future work, we did find diverse supporting evidence for each. As noted, the dimensions discovered are plausible and they explain much of the variance. In this we match best practice to date in the development and validation of dimensional models of style [33, 59, 60]. Further, we found that, although derived purely from prosodic features, many of the dimensions align with meaningful differences in lexical frequencies and with topic tendencies. Accordingly we are confident in our interpretations. Yet we also 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. It is certainly possible that there are more than 8 significant dimensions of style, but here our analysis of dimensions 9 and up did not converge on consistent interpretations. Perhaps a more sophisticated model would enable the discovery of more.

Incidentally, while there is no mathematical necessity for the dimensions derived from prosodic behavior frequencies to be informative regarding interaction styles, it turns out that they are. Also, there while there is also no mathematical necessity for the interpretations of each dimension’s poles to be opposed, for each dimension it turned out that they roughly are.

Further, while there is no necessity for the dimensions of style to have easy-to-grasp meanings, it turns out that they do, although admittedly to different extents. We might characterize them concisely as involving, respectively degree of: engagement, talker/listener role, positivity, knowledgeability role, speculation, fatalism, agreeableness, and emphaticness. However we caution that these terms are just mnemonics for the complex meanings presented above, and do not suffice to convey the dimensions’ full meanings.

We turn now to Research Question 3, regarding how previous descriptions of style aspects can be integrated into a general model.

Unfortunately direct comparisons are impossible, since no previous research released quantitative models. Nevertheless many previously-identified dimensions appear to align well with dimensions of this model. 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.
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.

We must also note that this set of 8 dimensions is only one of infinitely many sets of dimensions that span some of the space of variation. We do not claim that these 8 have any special status. Indeed, 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, though the resulting dimensions were always similar to these 8, they were never exactly the same.

Nevertheless, this set of 8 dimensions provides a new reference model, to which other observations on style tendencies and aspects can, if quantified, be mapped.

6 Illustrations of the Utility of the Model

Thus we have a model which can compute a reduced-dimensionality representation of the interaction style for any corpus, subcorpus, conversation, or conversation fragment. This section illustrates how this may be useful.

6.1 Corpus Characterization

This model supports visualization of data set differences. As an example, if we view Switchboard as a collection of subcorpora, one per topic, we can map out roughly where each topic lies by plotting the average interaction style of all fragments within that topic. Figure 5 is a projection of interaction style space. To avoid clutter, it shows only the 27 topics for which there was a lot of data, specifically 225 minutes or more, or which were among the 20 most distinctive topics, that is, those whose average interaction style was furthest from the global average.

While the aim of this model is not topic-based characterization, the positions of the topics in the figures suggest at least that it is picking up something meaningful. Further, we can see that it is not merely representing topic similarity by considering some of the topics that appear, at first glance, to be misplaced. For example, it may seems strange that conversations on the topic of “metric system” tend to be positive in style, but listening to examples shows that these conversations are mostly by engineers, who indeed discussed it positively. It may also seem strange that “woodworking” and “painting” are placed differently, as both can be at-home hobbies and projects. According to the model, their interaction styles are very different, averaging [0.6 2.2 -1.4 1.4 1.1 -0.4 0.6 0.5] and [-0.8 3.0 0.9 1.8 -0.6 0.1 -0.6 0.5], respectively, suggesting that dialogs about woodwork-

---

1 This subsection and the next are largely abbreviated from [91].
Figure 5: Average Interaction Styles of Some Topics in Switchboard, Projected to Interaction Style Dimensions 1 and 3. (0,0) is the global average style. The axis units are standard deviations computed over all conversation fragments. The topic names shown are just the mnemonics for the actual sentence-length prompts given to the participants.

...ing exhibited less shared engagement and were more positive and thoughtful in tone (Dimensions 1, 3, and 5, respectively). Listening confirmed that there was indeed a real difference, and it seemed that this was likely because the woodworking was mostly being discussed fondly by dedicated hobbyists, and the painting more often discussed by novices for whom it was a troublesome chore.

Thus the model provides an answer to Research Question 5: it can automatically characterize the style of any corpus, and thus help researchers and developers understand the diversity within and between corpora.
6.2 Similarity Estimation

Proximity in the model provides an easy to estimate similarity. For now, we can estimate similarity by simply using the Euclidean distance in the 8-dimensional space. For example, considering Switchboard’s 20 most distinctive topics, the closest pair was “politics,” at [1.0 2.6 0.1 1.6 0.4 0.1 0.1 -0.2] and “capital punishment,” at [1.1 2.7 0.3 1.6 0.5 0.0 0.0 -0.0]. The other most similar pairs were “baseball” and “football,” “weather/climate” and “vacation spots,” and “movies” and “TV programs.”

Such similarity estimates could be used to support targeted data augmentation. Considering again the scenario of seeking data to train a movie recommendation system, dialogs about movies have an average style of [-1.6 -0.0 0.5 0.0 -0.7 -0.5 .3 -.1], to which the most similar subcorpora were “TV programs” were “clothing and dress,” “football,” and “baseball.” This indicates what subcorpora would be potentially most compatible as supplementary data.

The similarity metric can of course be used not only to find close corpora or subcorpora, but also distant subcorpora or examples, for example to support data cleaning. For example, although many conversations in Switchboard have the “movies” tag, not all fragments are equally relevant. The model can identify atypical fragments as those most distant from the average interaction style. For the movies topic, examination of the five most distant fragments revealed that these were indeed mostly atypical — two involved strong moral judgments, and one was mostly about audience behavior — and would be good candidates for exclusion from the training set for a normal, upbeat movie recommending system.

If interaction-style similarity were no more than a reflection of topic similarity, its utility would be low, but these examples illustrate how it provides added value.

6.3 Exploring Demographic Differences

Sociolinguistics is largely concerned with differences across subpopulations. Differences in interaction styles are an issue of major practical importance, as they can impede successful communication. Classic work by Tannen and others has shown how people can misattribute style differences as problematic attitudes and intents [25, 26, 92]. 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 differences in interaction style have been hard to quantify, and our current understanding of this topic is largely based on anecdotal evidence. This model may be able to help. This subsection illustrates how.

The most straightforward use of the model is to compare average styles across subpopulations. Wanting to do a preliminary test of this method, but lacking data suitable for validation, we looked at two “subpopulations” for which the tendencies can be easily evaluated. In each Switchboard conversation, the first speaker to join the call was requested by the robot operator to “Please think about the topic while I locate another caller” [84], which sometimes took a few minutes. The later-joining speakers did not have this time to think. As seen in 8 the main differences between groups was that the former tended to talk more and to act more knowledgeable (Dimensions 2 and 4, effect sizes .04 and .05 standard deviations, respectively), as one would expect. (These differences were
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>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>
<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>
</tbody>
</table>

statistically significant (p < 0.0005, two-sided, unmatched-pairs, t-tests, with Bonferroni correction), but the effect sizes are small. These statements are also true for all findings reported in this subsection.

We then looked at gender differences. 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). The tendency to be more engaged aligns with cultural stereotypes, but the other two were unexpected. However we acknowledged that, since our model reflect prosodic behaviors only, this could be misleading, if, for example, the apparent tendency to negativity for women were balanced out by a tendency for them to be lexically more positive.

We also looked at age differences, comparing speakers older than 38 years old, the mean for this corpus, to those younger. Fragments with the older speakers tended to be more negative in style, and the older speakers tended to a more knowledgeable style (Dimensions 3 and 4, .13 and .10).

Table 9: 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.

<table>
<thead>
<tr>
<th>predictor</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>
<th>overall</th>
</tr>
</thead>
<tbody>
<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>
<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>
</tbody>
</table>

We also looked at the overall magnitude of the subpopulation differences. 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.

Crucially, our vector space representation enables us to measure the distance between any two interaction styles: we use the mean squared difference for each dimension, and also the Euclidean distance across dimensions. Of course this is useful only 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 seemed 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.

We use these distance measures in particular to evaluate prediction quality: the between a predicted style and the observed style is a measure of how informative the prediction is. As a baseline model of style predictability, we just predict the global average style for every fragment.

The first and second rows of Table 9 show how much better we do when predicting using two types of knowledge: the speaker’s gender and their age range. In each case, we predict the style of a fragment as the average of the interaction styles in fragments from that subpopulation. 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.

Further, 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.

These findings are only suggestive, not least because the Switchboard corpus was not collected to support comparisons across subgroup. Nevertheless, they illustrate how the model can be used to study population differences.

### 6.4 Exploring Individual Differences

A classic assumption in much user modeling work is that each individual has their own characteristic style. However, the validity of this assumption has rarely been quantified; previous work has treated only a few aspects of style, and has focused on whether individual tendencies exist, not how strong they are. Now, however, having a well-defined model of interaction styles, we are able to broadly examine the stability of individual styles (Research Question 4)\(^2\).

The first thing we note is that only two of the eight dimensions encode speaker-specific roles. The other six are collaborative or “aligned” [94], 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, representing who is the primary speaker and who is acting more knowledgeably, respectively, 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 each speaker’s fragments were located in a tight region of the space of styles. To quantify this, we again measured predictability. This is, inciden-
tally, consistent with Weise and Levitan’s [95] approach, and 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.

Again, as the baseline for each fragment we use the global style average. The predictors based on speaker identity instead predicted 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.

The last row of Table 9 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, contrary to the answer we expected for Research Question 4. This 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 that did not. In real life, we know that how people talk varies with the situation, topic, interlocutor, time of day, and other factors. But, given the quantity of research on accommodation and adaptation, we were surprised to see that individual styles count for so little.

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. For the most predictable speaker, the mean distance for predictions was only 50% of the average (she consistently took a passive listening role). However for the other 22%, predicting the speaker average was worse than predicting the global average. The most unpredictable speaker varied over 4 times the average. To infer what factors might be involved, we listening to some of her calls, and noticed that she had calling in at different times during the day, sometimes had a baby crying in the background, and had participated in conversations about a wide variety of topics.

Wondering whether perhaps accommodation simply takes time, we examined whether predictibility would would be easier for later fragments. Since entrainment in general takes time [96], 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 in fact the opposite was true: there was a slight tendency to less typical behavior over time.

Wondering whether the overall weakness of individual style consistency was obscuring a more detailed picture, we computed per-dimension reductions in prediction error, as seen in the last line of Table 9. We note that these were 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 infer that these
dimensions might be good candidates for aspects on which dialog systems might try to adapt to their user.

These findings are certainly not the last word on any issue in style stability and adaptation, but they do illustrate how this model can support systematic approaches to questions of practical importance.

7 Summary and Future Directions

The first contribution of this work is an expansion of our understanding of interaction styles. As summarized earlier in Table 3, we identified several previously-undescribed dimensions of style variation. Thus, thanks to the use of more features and more data, this work extends our understanding beyond the handful of style aspects identified by previous work.

The second contribution of this work is a workflow for discovering, from data, the dimensions of variation in interaction styles, together with the open-source software implementing it, available at https://github.com/nigelgward/istyles. Technically, the key innovation is in the feature set, which captures many aspects of dialog behavior.

The third contribution of this work is a few empirical findings: gender explains very little of the variation in interaction styles, the most stable aspect of speakers is their tendency to a positive vs. negative style, and individual styles explain little. The latter 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.

Finally, this work opens the path to deeper and broader examinations of interaction styles:

Future work should examine style 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 [97].

This model was built using only automatically computed features and simple mathematical procedures. While the result seems related to human perceptions of important aspects of style, and many of the perceived styles listed in Table 1 may possibly map to vectors or regions in this space, 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 most people are adept at subconsciously detecting and adopting various interaction styles, elucidating these abilities 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 [31].

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, based on Biber’s work suggesting that any reasonably diverse corpus can serve as a useful microcosm for all the genres of a language, it is likely that these 8 dimensions may have quite general utility. In any case, future work should seek to extend our understanding by applying these methods to larger, multi-genre data collections.

Future work should also examine connections between interaction styles and utterance-level style choice made by speech synthesizers and text generators. Recent work on these modules has sought to identify useful style control parameters, using both a priori design considerations and data-driven approaches, but always using decontextualized within-utterance data. It is possible that the new dimensions found here, based on analysis over longer time spans and exploiting prosodic information, could serve as additional, useful style control parameters.

Thus this model provides a starting place for important further investigations. Moreover, despite the many open questions, the model as-is may have value for support of corpus-based pragmatics research, for corpus selection, and for dialog system design and adaptation.

8 Acknowledgments

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

References


