Individual Interaction Styles: Evidence from a Spoken Chat Corpus

Nigel G. Ward
University of Texas at El Paso
nigelward@acm.org

Abstract

There is increasing interest in modeling style choices in dialog, for example for enabling dialog systems to adapt to their users. It is commonly assumed that each user has his or her own stable characteristics, but for interaction style the truth of this assumption has not been well examined. I investigated using a vector-space model of interaction styles, derived from the Switchboard corpus of telephone conversations and a broad set of prosodic-behavior features. While most individuals exhibited interaction style tendencies, these were generally far from stable, with a predictive model based on individual tendencies outperforming a speaker-independent model by only 3.6%. The tendencies were somewhat stronger for some speakers, including generally males, and for some dimensions of variation.

1 Introduction

To create dialog systems that are able to work very well for any user will require modeling and adapting to individual interaction styles (Eskenazi and Zhao, 2020; Marge et al., submitted, 2021). For example, Metcalf et al. (2019) demonstrated a Siri extension to detect which users are more talkative and then provide them information in a more chatty style. Sociolinguists, going back to Tannen (1980), have identified other ways in which people vary in interaction styles, such as focus on content *vs* interpersonal involvement, and domineering *vs* meek, among many others.

A general assumption, implicitly underlying much work across the broad area of user modeling and adaptation, is that that each user has consistent behavior tendencies. But how true is this for interaction styles? While variation and adaptation have been studied for many specific components — including utterance selection, lexical choice, speech synthesis, paralinguistic and turn-based prosody, and language generation (Eskenazi, 1993; Wang et al., 2018; Cao et al., 2020; Niu and Bansal, 2018; Hu et al., 2018; Cheng et al., 2019; Chaves and Gerosa, 2020) — the overall question seems not yet to have been examined. Thus this paper addresses, the questions of whether individual interaction styles exist and how much they explain. I also examine gender differences in style and adaptation, and other related questions.

2 Data

Work on individual differences in dialog has been limited, mostly using data sets with only a few dozen participants, and mostly considering only tightly structured dialogs, mostly task-oriented, but more speakers and more variety can lead to more general models. Most work has been limited to text or transcripts, but spoken data can be more informative. For these reasons I chose to use the Switchboard corpus of American English telephone conversations (Godfrey et al., 1992). Interaction styles are not instantaneous, but nor are they constant over long times, so I chose 30-second fragments as the unit of analysis. This seemed appropriate for a first study, and well-suited to Switchboard, where the topic, tone, and style often shift from minute to minute. Leaving some conversations for future validation work, I used a set of 33022 fragments, including 335 speakers.

3 Markers of Interaction Style

There are many possible choices for markers of interaction style. Like much previous work, I wanted to include prosodic features and features of turn-taking behavior (Grothendieck et al., 2011; Laskowski, 2014, 2016; Levitan, 2020), in part because being densely present, unlike word frequencies, they make analysis easier. However, wanting to consider more information, I created

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 27–31
July 29–31, 2021.

1	13%	both participants engaged	...	lack of shared engagement
2	11%	focal speaker mostly talking	...	focal speaker listening actively
3	8%	positive assessment	...	negative feelings
4	5%	focal speaker more dominant	...	nonfocal speaker more dominant
5	5%	factual	...	asking questions or speculating
6	4%	envisioning positive change	...	accepting things beyond individual control
7	3%	leading up to some larger point	...	making contrasts
8	3%	unfussed	...	emphatic

Table 1: Functions of the Top 8 Dimensions. The second column is the amounts of variance explained by each dimension, in terms of the 84 prosodic behavior frequency features.

a more inclusive set to track various prosodic behavior frequencies, including those relating to a wide range of dialog states, activities, and events, including many of those often considered most important in human interaction (Couper-Kuhlen and Selting, 2018), such as the extent and timing of turn holding, turn-taking, filler use and backchanneling; topic opening, development, and closing; bids for empathy; making positive and negative assessments; marking contrast; and so on. The specific features were based on a prosodic constructions model (Ward, 2019), in part because this enabled the use of a tool for automatic feature computation, including proper speaker and track normalization (Ward, 2021).

The feature computation starts by computing the quality of the match between each prosodic construction's prototypical configuration and the actual behavior of the interactants, every 20 milliseconds across each conversation fragment. Next, for each fragment, it computes the frequencies of occurrence for seven match-quality bins. For example, the fraction of timepoints at which the Enthusiastic Overlap Construction is strongly matching 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. Together these bin frequencies represent the extent to which the speakers are engaged in various interaction routines and the extent to which the dialog tends to dwell in certain states. With 12 prosodic configurations and 7 bins each, this gave 84 features per fragment.

4 The Space and the Dimensions

Given these 84 features, each fragment can be represented as a point in a 84-dimensional vector space. While hopeful that this space corresponds well with the perceptual space of interaction styles, for lack of previous work on perceptions of styles, I can here only present indirect evidence.

For current purposes, the most desirable property of this space is for fragments perceived closer in style to be closer in this space. Spot checking a few of the pairs that were closest in this space confirmed that each pair was indeed very similar in style.

Another desirable property is interpretability. Here, following Biber (2004), I choose to apply Principal Component Analysis to the data, expecting that the resulting dimension would be meaningful, thereby providing further evidence for the relevance of this space. Full discussion of the meanings of these dimensions will appear in another publication, but, in short, the top 8 dimensions indeed turned out to be meaningful, as revealed by good correlations with topics, lexical frequencies, and LIWC word categories frequencies. Table 1 summarizes. I illustrate the correlations seen by discussing Dimensions 3 and 6, chosen because there will later be interesting things to say about them.

One pole of Dimension 3 relates to a negative stance, with clear lexical tendencies: for example *gang, gangs, convicted, stole, offense*, and *disagree* all occurring over 3 times more commonly in these fragments. Topics in fragments near this pole were overwhelmingly things the speakers were not happy about, such as income tax, lawn problems, the futility of overseas aid, and time flying by. Prosodically, there is an overall lack of normal turn taking, with frequent long silences often serving to mark how breathtakingly inappropriate something was, for example the mathematical ignorance of junior college students, and frequent overlaps, often wryly sympathetic laughter. This style is also rich in the prosody of topic continuation and topic develop-

	dimension								distance
predictor	1	2	3	4	5	6	7	8	
speaker's average style	5.8%	4.0%	17.0%	2.5%	5.3%	8.0%	0.5%	2.7%	3.57%
gender average style	0.6%	0.0%	0.6%	0.0%	1.2%	0.1%	0.1%	0.4%	0.21%
age-range average style	0.1%	0.1%	0.3%	0.3%	0.0%	0.0%	0.4%	0.0%	0.06%

Table 2: Average prediction error reductions for various models: reductions per-dimension in mean squared error and reductions overall in Euclidean distance, all relative to always predicting the global average style.

ment, often used when piling up evidence for an opinion, for example about a politician. Conversely the other pole relates to a positive stance.

For Dimension 6, one pole involves a style of *accepting things beyond individual control*. This can involve situations like living in a small town where the big touring bands 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 relatively common words include *nope, uncomfortable*, and *weeds*. Conversely the other pole exhibits topic continuation prosody and a general lack of turn-taking, and relates to *envisioning positive change*.

Working in a reduced dimensionality space has numerous advantages, so for the analysis below I focused on just the top 8 dimensions. Checking the relationship between perceptual similarity and proximity in this simplified space, again by examining the closest pairs; again these were perceptually similar, and this was true in diverse regions of the space, 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 commercial policies that the interlocutor is also familiar with and views in the same way.

5 Measure and Models

Adaptive dialog systems need to predict what interaction style will be most appropriate for an upcoming dialog. Using speaker information should enable more accurate predictions, if indeed interaction styles are stable properties of individuals (Weise and Levitan, 2020). The vector space representation of styles enables us to measure the distance between any two interaction styles, and in particular, between a predicted style and the observed style. This can serve as a metric for the evaluation of predictive models of interaction style. Specifically, I use the mean squared difference for each dimension, and also the Euclidean distance across dimensions. While I report distance results below using only the top 8 dimensions, with all 84 the results were very similar.

The baseline model is to 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 for independent estimation of the individual's style. There were 31931 such fragments.

6 Results

The first row of Table 2 shows the reductions in prediction error obtained using the individual models, compared to the global-average baseline. Overall, knowing the speaker identity reduces the average prediction error by only 3.6%, a surprisingly modest amount.

However, predictability varied across speakers. Some were highly predictable: at one extreme, one speaker's mean distance for predictions was only 50% of the average (she consistently took a passive listening role); at the other extreme, one speaker's mean distance was over 4 times the average. Overall, speaker-specific knowledge enabled better predictions for 78% of the speakers.

Table 2 also shows the per-dimension prediction error reductions. The largest are 17% for Dimension 3, suggesting that for the negative *vs* positive dimension individuals tend to be relatively consistent, and 8% for Dimension 6, the resigned *vs*

progress-oriented dimension. Reductions for the other dimensions were all relatively low.

Digressing slightly, as entrainment in general takes time (Wynn and Borrie, 2020), one might expect that fragments taken from later into the calls would be closer to the participants' "true" styles, as they come to discover, reveal, relax into, and compromise towards their preferred styles. I therefore hypothesized that the styles of later fragments would be more predictable, but this turned out not to be the case.

7 Demographic Differences

The remaining rows of Table 2 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. Men and women are known to often differ significantly in interaction styles (Tannen, 1990), but here predictions based on gender are only about 0.2% better than generic predictions, and the age-class predictions show even less benefit. Thus, 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, I hypothesized that women would generally exhibit more style variation than men. 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 means had little predictive power, it is interesting to consider what the per-dimension tendencies suggest. I examined 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 more engaged, negative, and factual styles (Dimensions 1, 3, and 5, effect sizes .16, .16, and .22 standard deviations, respectively). Fragments with the older speakers tend to be more negative, and the older speakers tend to a more dominating style (Dimensions 3 and 4, .13 and .10). Fragments later in the conversation, specifically those occurring after 4 minutes in, tend to be more negative (.14). The speaker who joined the conversation first tended slightly to talk more and to dominate (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 et al., 1992), which sometimes took several minutes. All of these differences are statistically significant ($p < 0.0005$, two-sided, unmatched-pairs, t-tests with Bonferroni correction).

8 Discussion

While there was evidence that most individuals have their own interaction styles, these explained little, reducing the error of style predictions by only 3.6%. This implies that the styles are not very stable: that individuals vary greatly in style. Even if we could somehow create systems as good as the participants in this corpus at adapting their style to their interlocutor, they would generally perform only 3.6% better than systems that did not bother.

While this result came as a suprise to me, it is not really hard to understand; in real life we know that how people talk varies with the situation, topic, interlocutor, time of day, and other factors. This suggests that future research on interaction style adaptation for spoken dialog systems should prioritize adaptation to factors such as the topic, situation, and dialog activity type, rather than adaptation to the user.

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 interaction style is the extent to which the speaker tends to a positive or negative stance.

These findings and interpretations are tentative. Future work should examine the generality of this finding, with more features, various fragment sizes, more powerful models, and larger and more diverse data, including text-only dialogs. Future work should also examine not only behaviors but also preferences: although people in these conversations exhibited a variety of styles, perhaps, as users, people would prefer dialog systems that consistently use a fixed, individually-congenial interaction style. Examining this might further lead to a detailed understanding of preferences, leading ultimately to individualized mappings from system behavior to satisfaction properties (Yang et al., 2012). Finally, future work should include empirical explorations of human perception of the space of interaction styles.

To support such work, the code for the investigations so far is available at https://github.com/nigelgward/istyles.

9 Acknowledgments

I thank Aaron M. Alarcon for feature extraction code for a preliminary investigation, and Jonathan E. Avila, Olac Fuentes, and David Novick for discussion.

References

Douglas Biber. 2004. Conversation text types: A multi-dimensional analysis. In *Le poids des mots: Proceedings of the 7th International Conference on the Statistical Analysis of Textual Data*, pages 15–34. Presses Universitaires de Louvain.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. In *Association for Computational Linguistics, 58th Annual Meeting*, pages 1061–1071.

Ana Paula Chaves and Marco Aurelio Gerosa. 2020. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, 37:729–758.

Hao Cheng, Hao Fang, and Mari Ostendorf. 2019. A dynamic speaker model for conversational interactions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2772–2785.

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

Maxine Eskenazi. 1993. Trends in speaking styles research. In *Eurospeech*, pages 501–509.

Maxine Eskenazi and Tiancheng Zhao. 2020. Report from the NSF future directions workshop: Toward user-oriented agents: Research directions and challenges. *arXiv preprint arXiv:2006.06026*.

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

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

Zhichao Hu, Jean E. Fox Tree, and Marilyn Walker. 2018. Modeling linguistic and personality adaptation for natural language generation. In *Proceedings of the 19th annual SIGdial meeting on discourse and dialogue*, pages 20–31.

Kornel Laskowski. 2014. On the conversant-specificity of stochastic turn-taking models. In *Fifteenth Annual Conference of the International Speech Communication Association*, pages 2026–2030.

Kornel Laskowski. 2016. A framework for the automatic inference of stochastic turn-taking styles. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 202–211.

Rivka Levitan. 2020. Developing an integrated model of speech entrainment. In *IJCAI*, pages 5159 – 5163.

Matthew Marge, Carol Espy-Wilson, Nigel G. Ward, et al. submitted, 2021. Spoken language interaction with robots: Research issues and recommendations. *Computer Speech and Language*.

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

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

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

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

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

Nigel G. Ward. 2019. *Prosodic Pattterns in English Conversation*. Cambridge University Press.

Nigel G. Ward. 2021. Midlevel prosodic features toolkit (2016-2021). https://github.com/nigelg ward/midlevel.

Andreas Weise and Rivka Levitan. 2020. Decoupling entrainment from consistency using deep neural networks. ArXiv preprint arXiv:2011.01860.

Camille J. Wynn and Stephanie A Borrie. 2020. Classifying conversational entrainment of speech behavior: An updated framework and review. PsyArXiv.

Zhaojun Yang, Gina-Anne Levow, and Helen Meng. 2012. Predicting user satisfaction in spoken dialog system evaluation with collaborative filtering. *IEEE Journal of Selected Topics in Signal Processing*, 6:971–981.