Manipulating stance and involvement using collaborative tasks: An exploratory comparison

Valerie Freeman[1], *Julian Chan*[1], *Gina-Anne Levow*[1],
Richard Wright[1], *Mari Ostendorf*[2], *Victoria Zayats*[2]

[1]Department of Linguistics
[2]Department of Electrical Engineering
University of Washington
Seattle, WA USA
{valerief,jchan3,levow,rawright,ostendor,vzayats}@u.washington.edu

Abstract

The ATAROS project aims to identify acoustic signals of stance-taking in order to inform the development of automatic stance recognition in natural speech. Due to the typically low frequency of stance-taking in existing corpora that have been used to investigate related phenomena such as subjectivity, we are creating an audio corpus of unscripted conversations between dyads as they complete collaborative tasks designed to elicit a high density of stance-taking at increasing levels of involvement. To validate our experimental design and provide a preliminary assessment of the corpus, we examine a fully transcribed and time-aligned portion to compare the speaking styles in two tasks, one expected to elicit low involvement and weak stances, the other high involvement and strong stances. We find that although overall measures such as task duration and total word count do not indicate consistent differences across tasks, speakers do display significant differences in speaking style. Factors such as increases in speaking rate, turn length, and disfluencies from weak- to strong-stance tasks are consistent with increased involvement by the participants and provide evidence in support of the experimental design.

Index Terms: Stance-taking, involvement, conversational speech corpus, speaking style

1. Introduction

Stance-taking is an essential component of interactive collaboration, negotiation, and decision-making. When people take stances, they attempt to convey subjective feelings, attitudes, or opinions about the topic they are discussing [1, 2]. This can involve several levels of linguistic information, including acoustic, prosodic, lexical, and pragmatic factors. In theoretical linguistics, description of stance has been generally constrained to fine-grained content analysis, which often relies on subjective interpretation. An exception is the precursor to the current project [3, 4], which drew on existing frameworks in content analysis to identify areas for phonetic comparison. As some of the first work to focus on acoustic properties of stance-taking, it found that stance-expressing phrases had faster speaking rates, longer stressed vowels, and more expanded vowel spaces when compared to stance-neutral phrases. Such acoustically-measurable properties are the target of investigation in the ATAROS (Automatic Tagging and Recognition of Stance) project, which includes the collection of the stance-rich audio corpus described below. Collecting the corpus will allow us to identify and quantify properties of the speech signal associated with stance-taking, create an acoustic model of stance, and test theories of stance-taking on natural speech.

In automatic recognition research, stance links most closely to sentiment and subjectivity, expressions of a “private state” [5], an internal mental or emotional state. Research on sentiment and subjectivity analysis has exploded since the publication of foundational work such as [6, 7]. The majority of this work has focused on textual materials with accompanying annotated corpora, such as those described in [7, 8, 6] and many others. Such text-based approaches to subjectivity recognition primarily exploit lexical and syntactic evidence, relying on long, well-formed sentences and clauses for identification of stance-taking. However, our focus is on stance-taking in spoken interactions, which involve short, fragmentary, or disfluent utterances. Importantly, conversational speech harnesses information beyond its textual content to convey information about stance, for example through intonation, speaking rate, stress, and precision of articulation [3, 4]. In general, issues of subjectivity, sentiment, and stance in speech have received much less attention, and this work has primarily relied on existing conversational dyadic ([9] in [10]) or multi-party meeting data ([11, 12, 13] in [14, 15, 16], respectively). In these cases, a small portion of the existing corpus was annotated for subjectivity or related factors such as agreement or arguing. Even using speech data, many of the approaches to automatic subjectivity recognition have relied primarily on word or n-gram content [17], and their efforts to incorporate prosodic information yielded no significant improvement [18]. However, [15] found that access to the additional information in the audio channel of meeting recordings enabled annotators to better identify opinions, especially negative opinions, than using transcripts alone. Since the understanding of the factors of speech which convey stance-related information is still in its early stages, we employ a bottom-up strategy of creating a corpus to elicit varying levels of stance-taking and involvement and analyzing differences in speaking style across these conditions.

The ATAROS corpus is designed specifically for the purpose of identifying acoustically-measurable signals of stance-taking in natural speech, and as such, it provides several advantages over speech collected for other purposes. Limitations of existing corpora include issues with recording quality, speaker attributes, and the type and content of speech. Recording quality varies widely when audio is gathered from sources not created for linguistic analysis. Common concerns are recording conditions and microphone type and placement, which often affect the signal-to-noise ratio and acoustic intensity. For example, if the distance between speaker and microphone varies unpredictably, intensity is an unreliable measure, just as it is when loudness is adjusted for public broadcast (TV, radio, etc.).

More specific to the study of linguistic variation is the ability to disentangle within- and between-speaker variation. Factors to consider include speaker demographics, social roles, and the amount and type of speech collected from each person. Social factors such as age, gender, ethnicity, dialect region, and the relationship between speaker and audience commonly correlate with linguistic variation [19, 20, 21], but these attributes are not always known or controlled in audio collections. This was a problem in the precursor to the ATAROS project [3, 4], in which the political talk show under analysis contained only males, each from a different dialect region (which was only possible to determine because they happened to be reasonably well-known people with publically-available biographic information). The type of speech also matters; of interest here is stance in spontaneous, unscripted, naturalistic conversation, which differs from read or performed speech in ways that may affect stance-taking. For example, the personal motives underlying stance moves may differ greatly between social roles (boss, friend, parent, etc.) and between settings (meetings, public discussion, personal conversation, etc.). More to the point, many situations do not naturally involve a high density of the phenomenon under investigation. This is particularly relevant for stance-taking, which might be found in high densities in more formal, scripted situations such as debates but less reliably in conversation. Finally, when intra-speaker variation is desired, a larger amount of speech is required from each speaker in each condition predicted to have an effect, in order to obtain enough power for linguistic analysis and to provide sufficient material for computational modeling and machine learning.

All of the above factors are addressed in the ATAROS corpus. Its high-quality audio recordings are ideal for acoustic analysis, with head-mounted microphones in separate channels and a quiet environment. Conversation is unscripted but focused around collaborative tasks that require increasing levels of involvement and stance-taking. With some structure provided by the tasks, many target words are repeated throughout the recordings, enabling straightforward comparisons within and across both speakers and tasks. All speakers complete all tasks in one session, yielding a similar amount of speech in each task from each speaker. Basic demographics are collected and controlled: speakers are matched roughly for age and either matched or crossed by gender, yielding approximately equal numbers of male-male, female-female, and male-female dyads. All are native English-speakers from only one dialect region, the Pacific Northwest. Controlling for dialect region is especially useful in these initial stages of isolating linguistic behavior attributable to stance or involvement without the potential confound of differences between dialects (e.g., vowel inventories, pause durations, pitch patterns, backchannel behavior; [3]).

There is a natural tension between the use of a carefully controlled dataset for analysis and the desire to apply such analysis to the growing wealth of naturally occurring spoken language. The former allows greater control and power, bringing subtle contrasts into sharp relief. The latter allows evaluation on the types of data in real-world applications, where such controls are impossible. Thus we plan to collect a well-controlled corpus of stance-taking to fully explore the range of associated linguistic behaviors and then to generalize these findings by applying them to a larger-scale, high stakes, naturalistic corpus, the Congressional Hearings of the Financial Crisis Inquiry Commission.

A:	Books could go near toys I think. Maybe.
B:	Yeah or travel guide- Yeah, between toys and travel guides?
A:	Yeah, sure.

Table 1: Snippet from the Inventory Task, designed to elicit low involvement and weak stances.

The rest of the paper is organized as follows. In Section 2, we describe our corpus collection and transcription procedures, including details on the design of the two collaborative tasks for which we compare descriptive measures of speaking style in Section 3. Finally, Section 4 summarizes our current findings and our plans for stance annotation.

2. Corpus Design, Collection, Transcription

In this section, we present the tasks designed to elicit varying degrees of stance-taking and involvement, the data collection process, and transcription and alignment methodology.

2.1. Task Design

Each dyad completes five collaborative problem-solving tasks designed to elicit frequent changes in stance and differing levels of involvement with the task. There are two groups of tasks, each of which uses a set of about 50 target items chosen to represent the main vowel categories of Western American English in fairly neutral consonantal contexts (i.e., avoiding liquids and following nasals, which commonly neutralize vowel contrasts). Each group of tasks begins with a find-the-difference list task intended to elicit stance-neutral first-mentions of the items to be used in subsequent tasks. The other three tasks are designed to elicit increasing levels of involvement and stronger stances. In this report, we will describe and compare the low- and high-involvement tasks, Inventory and Budget.

The *Inventory Task* is a collaborative decision-making task designed to elicit low levels of involvement and weak stances. Speakers stand facing a felt-covered wall and are given a box of about 50 Velcro-backed cards that can be stuck to the felt. The cards are printed with the names of household items, and about 15 additional cards are already placed on the wall, which represents a store inventory map. Speakers are told to imagine that they are co-managers of a superstore in charge of arranging new inventory. Their job is to discuss each item in the box and agree on where to place it; once it is on the wall, it cannot be moved. This task generally involves polite solicitation and acceptance of suggestions, as seen in the excerpt in Table 1.

The *Budget Task* is a collaborative decision-making task designed to elicit high levels of involvement and strong stances. Speakers are seated at a computer screen and told to imagine that they are on a county budget committee in charge of making cuts to four departments. About 50 services and expenses are divided among the four departments on the screen. Their job is to discuss each item and decide whether to fund or cut it; the only limitation is that they must cut the same number of items from each department. This task involves more elaborate negotiation, which may include citing personal knowledge or experience as support for stances. An example of this appears in the excerpt in Table 2.

A:	Well job training programs is pretty crucial. [...] And so is .. chicken pox vaccinations, right?
B:	I - well, I didn't get a chicken pox vaccination. I think a lot of kids just naturally get chicken pox and then they're fine.

Table 2: Snippet from the Budget Task, designed to elicit high involvement and strong stances.

2.2. Recording conditions

Recordings are made in a sound-attenuated booth on the University of Washington campus in Seattle. The booth measures approximately 7 feet by 10 feet and contains a card table, 2-4 chairs, and a small heavy table with a computer screen and keyboard. Each participant is fitted with a head-mounted AKG C520 condenser microphone connected by XLR cable to a separate channel in an M-Audio Profire 610 mixer outside the booth. The mixer is connected to an iMac workstation that uses Sound Studio (version 3.5.7) to create 16-bit stereo WAV-file recordings at a 44.1 kHz sampling rate. The computer screen in the booth is connected to the iMac as a second monitor where instructions are displayed for two of the tasks.

2.3. Speakers and corpus size

Speakers are native English-speakers age 18-75 who grew up in the Pacific Northwest Washington, Oregon, Idaho). Of the 26 dyads recorded so far, 5 are male-male, 9 female-female, and 12 mixed-gender, for a total of 22 males and 30 females. About half of these are under age 30 (11 males, 14 females), a quarter are 30s through 40s (7 males, 6 females), and a quarter are over 60 (4 males, 10 females). Recording will continue toward the goal of at least 30 dyads divided evenly by gender condition and age group. Total recording time for all five tasks combined normally ranges from 40 to 80 mins per dyad. With an average of about 60 mins per dyad, the corpus is expected to yield at least 30 hours of dyadic interaction. Currently, all dyads are strangers matched roughly by age, but future recordings may include pairs of friends or combinations of age groups.

2.4. Transcription

Tasks are manually transcribed at the utterance level in Praat [22] following a simplified version of the ICSI Meeting Corpus transcription guidelines [11]. Stretches of speech are demarked when surrounded by at least 500 ms of silence; pauses shorter than 500 ms are marked within an utterance with two periods. Every word is transcribed orthographically using conventional American spelling, with the addition of common shortenings (cuz, kay, etc.), phonological contractions (gonna, wanna, hafta; kinda, sorta, etc.), discourse markers (uh-oh, mm-hm, etc.), and vocalizations with recognized meanings (e.g., psst, shh; meh (verbal shrug), psh (verbal scoff)). Filled pauses are transcribed as "uh" or "um," with the latter indicating audible nasality. Disfluencies are marked with a short dash, without a space for truncated words (e.g., categ-) or following a space for uncompleted thoughts (e.g., I thought -), which may end an utterance or precede a repetition or restart (e.g., I don't - I'm not - I'm not sure.). A small, finite set of vocalizations are transcribed with tags (e.g., {VOC laugh}, {VOC cough}), and notable voice qualities or unusual pronunciations are marked with a following descriptive tag (e.g., {QUAL laughing}). Utterances are transcribed using conventional capitalization and a limited set of punctuation marks, e.g., period to end a complete statement, question mark to end a syntactic question, commas to separate lists (no colons, semi-colons, or quotation marks are used).

	Inventory	Budget
# Dyads	12	12
# M-M	3	3
# F-M	6	6
# F-F	3	3
Total Duration	2h 24m	2h 14m
Ave. Duration	≈12m (2.25m)	≈11.2m (5m)
Total Trans. Wds	20468	21887
Ave. Trans. Wds	1705	1824
Total Turns	3527	3104
Ave. Turns	294	259

Table 3: Overview of the Inventory and Budget Tasks, by speakers, duration, words, and turns. Standard deviations appear in parentheses.

Completed manual transcriptions are automatically force-aligned using the Penn Phonetics Lab Forced Aligner (P2FA; [23]), which demarks word and phone boundaries. Transcribed words not already in the pronouncing dictionary provided with P2FA (CMUdict v. 0.6) (place names, truncations, vocalizations, etc.) are added as needed.

3. Preliminary Corpus Analysis

In order to begin modeling behavior with differences in stance and involvement, the Inventory and Budget tasks are prioritized for transcription and annotation, as they are expected to yield the lowest and highest levels of stance and involvement. To date, the Inventory and Budget tasks for 12 dyads have been transcribed and force-aligned to the audio signal. This subset allows us to begin to characterize the corpus being collected, to assess the differences in speech across the different task settings, and to investigate the contrasts in speaking style associated with differing degrees of stance-taking. Table 3 shows a broad description of the recordings collected for these 24 tasks in terms of total and average task duration and words transcribed, as well as speaker distribution.

It is clear from the overview in Table 3 that the total time and activity spent on each of the tasks is quite similar overall, in spite of their anticipated differences in levels of stance-taking. They also exhibit substantial variability: although the mean Inventory task duration is 12 minutes, the standard deviation is 2.25 minutes. The Budget task, in turn, has mean duration of just over 11 minutes as a standard deviation of over 5 minutes.

To understand whether there are systematic differences in speaking style that are being elicited by these different task settings, we explored several measures that can be directly extracted from the time-aligned transcripts. Specifically, we compared overall speaking rate (in syllables per second), turn duration (in transcribed words), and disfluency rates based on two different measures described in more detail below.

We conducted within-speaker comparisons for all measures across the two task conditions based on Wilcoxon signed-rank test. There is a marked difference in speaking style across the two task conditions, with a significantly faster speaking rate ($p < 0.001$) and with average turn length greater in the Budget task than the Inventory task ($p < 0.001$). This increase in turn length is illustrated in Figure 1, which presents histograms

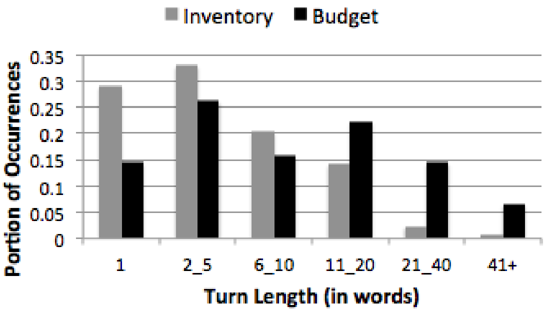

Figure 1: Contrast in turn lengths in words between weak-stance Inventory task (grey) and strong-stance Budget task (black) for a male example speaker. Lengths are binned according to the ranges on the x-axis.

of the turn length (in words) for an examplar male speaker under the two task conditions. While the Inventory task (shown in grey) exhibits a particularly high rate of single-word utterances (typically backchannels), the Budget task (shown in black) displays a much lower rate. Further, the Inventory task turns are concentrated in the shorter durations, while those in the Budget task are longer overall. While these measures were computed using manually transcribed utterance boundaries, a comparable analysis using automatically detected segments based on silences from the forced alignment yielded analogous patterns of significantly longer speech segments and significantly faster speaking rates.

Multiple types of disfluencies are observed in spontaneous speech, including filled pauses, repetition disfluencies, restarts, and correction repairs. While some disfluencies can be difficult to automatically detect, repetition disfluencies can be detected in multiple genres with relatively high accuracy using a model trained on the Switchboard corpus, e.g. F-measures of roughly 0.8-0.9 [24, 25]. Thus, it is possible to use automatic annotation in comparing some disfluency differences across conditions. We find a significant difference in the automatically annotated repetition rate between the weak- and strong-stance conditions ($p < 0.01$), with the Budget task characterized by repetition rates 68% higher than those observed in the Inventory task.

Our manual transcription also enables an additional investigation of disfluency rates. Specifically, the corpus annotation includes marking of filled pauses, "um" and "uh" distinguished by the presence or absence of nasality, as well as labeling of truncated words. We compute the combinated rate of these labeled disfluencies per transcribed speaking turn. Overall, speakers are significantly more disfluent in the Budget task than the Inventory task by this measure ($p < 0.05$). Male speakers are particularly disfluent and exhibit an average increase of 34% in the Budget task over that in the Inventory task. These findings - faster speech, longer utterances, increased disfluencies - are consistent with higher levels of involvement, as intended in our task design.

4. Conclusions & Future Work

We have described the motivation and design for the collection of the ATAROS corpus, a corpus of dyadic conversational speech focused on eliciting varying levels of stance-taking and involvement. Focusing on two tasks targeting extremes of weak and strong stance-taking allows assessment of this protocol and a preliminary investigation of the speech produced under these conditions. The task designed to elicit stronger stances and greater involvement exhibits longer turns and increased rates of disfluency, as measured by both manually labeled filled pauses and truncated words and automatically detected repetitions. These behaviors are consistent with increased levels of involvement in the conversation and provide evidence for the effectiveness of the experimental design.

An initial release of the corpus is available for research purposes through the UW Linguistic Phonetics Lab website (`http://depts.washington.edu/phonlab/projects/ATAROS`). Future releases will include both coarse- and fine-grained stance annotation. At the coarse level, we annotate spurts holistically for degree of stance-taking (no stance, weak stance, moderate stance, strong stance) and also for polarity (positive or negative). At the fine-grained level, we tag word and word sequences specifically indicative of stance-taking, using the scheme developed in [3, 4] and extended and validated in a pilot study for the current corpus. This scheme grounds stance annotation in lexical and semantic content, drawing on an array of indicators, such as overt evaluation, modifiers and intensifiers, citing of evidence, negotiation, agreement, and disagreement. The resulting corpus will thus support the in-depth investigation and acoustic-phonetic characterization of the linguistic expression of stance-taking in conversational speech.

5. Acknowledgements

Creation, annotation, and analysis of this corpus is supported by NSF IIS: #1351034. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF or the U.S. Government. We would also like to thank the transcriber/annotators who have contributed to this work: Heather Morrison and Lauren Fox.

6. References

[1] D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan, *Longman grammar of spoken and written English*. Longman, 1999.

[2] P. Haddington, "Stance taking in news interviews," *SKY Journal of Linguistics*, vol. 17, pp. 101–142, 2004.

[3] V. Freeman, "Using acoustic measures of hyperarticulation to quantify novelty and evaluation in a corpus of political talk shows," Master's thesis, University of Washington, 2010.

[4] ——, "Hyperarticulation as a signal of stance," *Journal of Phonetics*, vol. 45, pp. 1–11, 2014.

[5] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik, *A Comprehensive Grammar of the English Language*. New York: Longman, 1985.

[6] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002, pp. 79–86.

[7] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, vol. 39, no. 2–3, pp. 165–210, 2005.

[8] S. Somasundaran and J. Wiebe, "Recognizing stances in online debates," in *Proceedings of ACL 2009: Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, 2009.

[9] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proceedings of ICASSP-92*, 1992, pp. 517–520.

[10] G. Murray and G. Carenini, "Detecting subjectivity in multiparty speech," in *Proceedings of Interspeech 2009*, 2009, pp. 2007–2010.

[11] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at ICSI," in *Proceedings of Human Language Technologies Conference*, 2001.

[12] S. Burger, V. MacLaren, and H. Yu, "The ISL meeting corpus: The impact of meeting type on speech type," in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 2002.

[13] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus," in *Proceedings of the Measuring Behavior Symposium on "Annotating and Measuring Meeting Behavior"*, 2005.

[14] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement vs. disagreement in meetings: Training with unlabeled data," in *Proceedings of HLT-NAACL Conference*, Edmonton, Canada, 2003.

[15] S. Somasundaran, J. Wiebe, P. Hoffmann, and D. Litman, "Manual annotation of opinion categories in meetings," in *ACL Workshop: Frontiers in Linguistically Annotated Corpora(Coling/ACL 2006)*, 2006.

[16] T. Wilson, "Annotating subjective content in meetings," in *Proceedings of the Language Resources and Evaluation Conference*, 2008.

[17] T. Wilson and S. Raaijmakers, "Comparing word, character, and phoneme n-grams for subjective utterance recognition," in *Proceedings of Interspeech 2008*, 2008.

[18] S. Raaijmakers, K. Truong, and T. Wilson, "Multimodal subjectivity analysis of multiparty conversation," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, October 2008, pp. 466–474. [Online]. Available: http://www.aclweb.org/anthology/D08-1049

[19] W. Labov, "Social motivation of a sound change," *Word*, vol. 19, pp. 273–303, 1963.

[20] S. R. Fussell and R. M. Krauss, "Understanding friends and strangers: The effects of audience design on message comprehension," *European Journal of Social Psychology*, vol. 19, no. 6, pp. 509–525, 1989.

[21] H. J. Ladegaard, "Audience design revisited: Persons, roles and power relations in speech interactions," *Language and Communication*, vol. 15, no. 1, pp. 89–101, 1995.

[22] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]. version 5.3.55," 2013, http://www.praat.org.

[23] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," in *Proceedings of Acoustics '08*, 2008.

[24] V. Zayats, M. Ostendorf, and H. Hajishirzi, "Multidomain disfluency and repair detection," in *Proceedings of Interspeech*, 2014.

[25] M. Ostendorf and S. Hahn, "A sequential repetition model for improved disfluency detection," in *Proceedings of Interspeech*, 2013.

[26] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, M. Strube and C. Sidner, Eds. Cambridge, Massachusetts, USA: Association for Computational Linguistics, April 30 - May 1 2004, pp. 97–100.