A voice is worth a thousand words: the implications of the micro-coding of social signals in speech for trust research

Benjamin Waber, MIT Media Laboratory

Michele Williams, Cornell University

John Carroll, Sloan School of Management

Alex "Sandy" Pentland, MIT Media Laboratory

Summary

While self-report measures are often highly reliable for field research on trust (Mayer and Davis, 1999), subjects often cannot complete surveys during real time interactions. In contrast, the social signals that are embedded in the non-linguistic elements of conversations can be captured in real time and extracted with the assistance of computer coding. This chapter seeks to understand how computer-coded social signals are related to interpersonal trust.

Introduction

Self-report measures of trust reflect an important and often highly reliable tool for researchers interested in trust (Mayer and Davis, 1999). However, self-report measures require subjects to stop and think about how much they trust others or are trusted by others. Researchers are not able to use these methods when subjects cannot stop to fill out surveys in real time. In our setting, medical conversations or handoffs, one member of the pair must quickly receive critical information about a patient's current medical condition and then immediately begin caring for that patient. The rushed and technical nature of these conversations also makes qualitative

research difficult because most of the social signals embedded in these conversations are non-verbal. During a transition in care, such as those we observed, medical personnel rarely stopped to relay social information verbally, making transcripts of their conversations useless for retrieving social content. Although video recording and coding of non-verbal behavior such as eye contact is an option, it is more invasive than audio recording.

The social signals embedded in the non-linguistic elements of conversation reflect a source of relational information that has received little research attention from trust scholars (Curhan & Pentland, 2007; Pentland, 2004). Thus in this chapter, we seek to understand how the social signals embedded in non-linguistic elements of conversations are related to interpersonal trust. Non-linguistic elements of conversations include voiced utterances, which are vowel sounds like /o/, and unvoiced utterances, which are everything else such as bursts from the lips like /p/ and fricatives like /s/. They also include features of one's voice such as emphasis. Emphasis, for example, is determined by both the loudness of your voice and its pitch, i.e., how high like a soprano or low like a baritone your voice sounds. The non-linguistic elements of a conversation exclude "content;" that is, the information or meaning contained in the words or sentences you utter.

Specifically, our research team studied the social side of technical communication in a major hospital setting. Although effective communication during transitions in care is known to be essential for the continuity of patient care in hospitals, we only have a partial understanding of the interpersonal communication behaviors that health care providers can use to enhance both the accurate transfer of clinical information and the maintenance of interpersonal trust.

Our research group faced the challenge of understanding and measuring the non-linguistic elements of technical medical conversations. Because these conversations focused on patients' medical conditions, our transcripts provided little insight into the relational elements of the conversation. However, relational aspects of these conversations were present in the non-linguistic elements of the participants' speech, such as activity level and emphasis (a measure which combines pitch and loudness). These elements reflect aspects of people's engagement in the conversation and their relational responsiveness to one another. We argue that non-linguistic elements of the participants' speech not only influence the transfer of technical information, but also affect participants' experience of trust during these interactions. In this chapter, we will describe the challenges and benefits of using computers to code the non-linguistic elements of conversations and present results from a pilot study.

Description of the methods

In this chapter, we investigate the use of computer coding of non-linguistic aspects of speaking patterns such as emphasis and activity level. Humans and to a certain extent all mammals have evolved to pick up on "social signals." That is, they interpret the non-verbal behavior of others, such as non-linguistic elements of speech (Pentland, Honest Signals, 2008). Even watching a foreign film when you don't understand the language, you can still pick up the gist of what's occurring: which characters are interested in the conversation, who is positioning themselves in a dominant role, etc. Similarly, you can infer what your dog or your baby is feeling, not from what they say, but the way that they vocalize or move.

Psychologists recently began to take an interest in these signals, particularly after the groundbreaking research by Ambady and colleagues (Ambady, LaPlante, Nguyen, Rosenthal, Chaumenton, & Levinson, 2002). Subjects in this study were asked to listen to 20 seconds of a doctor patient conversation and then judge whether or not the patient would sue the doctor for malpractice. They found that subjects that listened to the audio with the content (i.e., audible words) filtered out did *as well* as subjects who heard the content.

Despite this research, coding the non-linguistic elements of conversations from audio and video recordings typically presents the challenge of agreement among multiple human raters. In addition, certain features of speech such as speaking speed simply cannot be coded without computational aids. Thus, in this chapter, we investigate the use of computer coding of non-linguistic aspects of speaking patterns such as emphasis.

Sumit Basu and Alex Pentland at the MIT Media Lab began developing methods in 2001 to automatically segment human speech and extract useful features (Basu, 2002; Pentland, 2004). Although coding the non-linguistic elements of conversations from audio and video recordings presents challenges, Pentland (2004) constructed measures for four types of vocal social signaling: activity level, engagement, emphasis, and mirroring. These four measures were extrapolated from a broad reading of the voice analysis and social science literature.

Activity level refers to how long a person is speaking combined with how much emphasis is present in their speech. It is computed by breaking up speech into speaking and non-speaking segments and computing the voiced segments to determine emphasis. Engagement refers to how much each person is influencing the pace of the conversation, and it is determined by examining how the average pace of a conversation for an individual is changed in the current conversation.

Emphasis as described earlier combined information about the loudness and pitch of one voice and reflects the amount of deviation in these values from their mean. Socially, emphasis indicates the importance that the speaker puts on an interaction (Curhan and Pentland, 2007). Finally, mirroring, which may signal empathy, is defined as the amount of short interjections uttered by both conversational participants (for example: "OK? OK!") (Curhan and Pentland, 2007). (See Appendix A for a technical description of these measures.)

By modeling the way humans produce speech, Pentland and colleagues were able to achieve unprecedented accuracy in speech segmentation and created new, compelling models of conversational dominance. Pentland and colleagues are now working to establish the general validity and nomological networks of these measures.

Recently, Pentland's group applied these techniques in experimental settings. One study examined the voice features of two people involved in a salary negotiation. By calculating four simple voice features, such as activity level, conversational engagement, prosodic emphasis, and vocal mirroring, the authors were able to predict 30% of the variance in the final salaries ((i.e., $R^2 = 0.30$ in the regression equation, Curhan & Pentland, 2007). Although not measured in this study, it stands to reason that because vocal features reflect one's interest in one's conversational partner, they would not only influence the substantive outcome of a negotiation but also the relational outcome. For example, the trust and relationship quality established during a negotiation.

Madan, Caneel, and Pentland (2004) showed the generality of these features by performing a similar experiment in a speed dating scenario. In this experiment, pairs of one man and one woman sat at separate tables and talked for 5 minutes with the purpose of determining

whether or not they should go out on a date. After 5 minutes the males changed tables and talked to a different female. At the end of the exercise, each person rated their interest in dating the people they talked to. Collecting the same features as those mentioned above, the authors were able to predict 40% of the variance in responses (i.e., $R^2 = 0.44$ in the regression equation). Although not measured, trust is likely to be an important component of people's interest in dating one another and thus predicted by vocal features.

Our research seeks to establish the link between vocal features of a conversation and interpersonal trust. This line of study has the potential to enable future research using vocal features as (1) a proxy for trust or relationship quality, (2) an antecedent of trust building and (3) a moderator of positive and negative vocal content (the meaning of a conversation) on trust.

In this chapter, we focus on the non-linguistic feature of emphasis (pitch and volume), which had a consistent relationship with both trust and information transfer.

Personal experience of micro-coding of speech

Trust facilitates information sharing and knowledge transfer (Currall and Judge, 1995; Levin and Cross, 2004). Trust enhances self-disclosure and allows individuals to ask questions without the fear of being taken advantage of (Levin and Cross, 2004). However, little is known about the relationship between trust and non-linguistic speech behavior. We argue that the non-linguistic components of speech carry the relational content of technical conversations and also influence the effectiveness of speech. For instance, emphasis (a combined measure of pitch and volume) indicates the importance that the speaker puts on an interaction (Curhan and Pentland, 2007) and also focuses listener attention on specific content or information that the speaker

believes is most important. We argue that trust increases the use of relational speech features and that relational speech features, in turn, should enhance both the transfer of technical information and subsequent trust.

In our study of medical transitions in care, we studied nurses from a large urban hospital. These nurses were engaged in transferring the information required for the ongoing care of actual patients. We recorded the specific constrained interaction situations in which one outgoing nurse transferred the medical information associated with a patient to an incoming nurse who would then care for the patient during the upcoming shift. Our data was dyadic by handoff. Using a sample of 29 nurses in 45 unique dyads and a fixed effects model, we used computer coding of speaker dyads to investigate the impact of non-linguistic features of communication on the transfer of technical information and interpersonal trust. The raw audio files from the interactions were fed into a computer program, which then performed voicing analysis and speaker identification. Next higher-level features such as loudness and pitch were computed and used to create the activity, engagement, emphasis, and mirroring features.

We found that emphasis (i.e., variations in pitch and volume) that partners used in their speech mattered. Emphasis, which reflects emotional engagement in a conversation, was significantly associated with the technical adequacy of the information transfer as coded by an independent nursing expert, but it was not related to the trust experienced during the transfer as reported by the dyad partners. In an additional, individual level analysis, however, a nurse's trust in his or her colleagues measured several weeks prior to the observed transition in care [1] was significantly related to the variation in emphasis used by that same outgoing nurse during the

[1] Trust had been measured earlier as the psychological safety of the outgoing nurse (i.e., willingness to trust or make oneself vulnerable to the other nurses on the unit, Edmondson, 1999).

observed transition. Thus, our preliminary findings suggest that trust may form a context that influences the use of non-linguistic elements of conversation, elements which in turn are related to the accuracy of information transfer.

Research validity and caveats: challenges for recording and coding trustful conversations

Recent methods for voice analysis have been developed through computationally modeling the speech production process (i.e. how air is compressed in the vocal cords and modified by the tongue), as well as extensive training of the data processing software on large data sets to determine appropriate settings for determining speaking/non-speaking and voiced/unvoiced thresholds. These methods are also robust to noise and microphone distance. In particular, under outdoor settings researchers have correctly labeled 98.3% of the voicing information as well as 99.3% of the speaking information (Basu, 2002). Here voicing information can be thought of loosely as vowel sounds or voiced utterances and unvoiced utterances, which are everything else such as bursts from the lips like /p/ and fricatives like /s/.

However, recording voice data is challenging in a dynamic setting such as medical transitions in care, where dyads are talking in a crowded room. For instance, our data collection consisted of direct observation and audio taping that took place in the nurses' lounge of a 30-bed medical surgical unit of a large urban teaching hospital. The room was approximately twelve by twelve feet square with a large round table in the middle. The room also had a refrigerator, microwave, toaster oven, and various cabinets. Depending on how many nurses were going off and on shift, the room would have from four to ten nurses who would all be speaking in dyads at the same time.

A difficult problem with such unconstrained contexts is detecting who is interacting with whom. Recently, conversations have been detected and isolated with reasonable success (Choudhury, 2004). Even when conversation detection is accurate, however, the corresponding audio features lose some of their predictive power in unconstrained settings (Wu, Waber, Aral, Brynjolfsson, & Pentland, 2008). This is most likely due to the fact that the topics of conversations can vary widely in these situations, making it more difficult to isolate speech patterns related to work versus purely social conversations.

In our setting, not only did each participant in the conversation need to wear a recording device, but in addition, the placement of each device had to ensure that both voices in the conversation were not equidistant from any one recording device. In our pilot study, we found that only 45/70 of the recordings (unique and repeating dyads) had sufficient quality for the computer to easily extract vocal features. The remaining recordings required human intervention to process them accurately. Significant human intervention was also necessary for preprocessing the recordings in order to extract features from the audio data.

Discussion: implications for organizational research

Our pilot data suggests a link between trust and non-linguistic features of speech that, in turn, enhance the transfer of technical information. In our medical setting, this enhanced information transfer has implications for patient safety. For example, communication breakdowns were considered to be the primary root cause of over 60% of the sentinel events in a national sample

of preventable errors in hospitals.[2] At our research site, communication breakdowns were identified as a contributing factor in 31% of the asserted malpractice claims. Thus, because effective communication has implications for safety, the relationship among trust, non-linguistic features of speech, and effective communication may be important for a variety of high reliability organizations.

We were surprised that the non-linguistic features of speech were not significantly related to our measures of trust during the observed transition in care, but only to trust measured several weeks prior to the observed transition. However, we believe that the survey that nurses filled out after their interaction may have been compromised by the fact the outgoing nurses were rushing to go home and incoming nurses were rushing to see their patients. Another "real-time" measure of relationship quality such as the physical proximity between dyad partners during the interaction may help reveal the more relational implications of the non-linguistic coding. Alternatively, it may also be the case that trust in this situation is related to competence (i.e., the quality of the information provided by the outgoing nurse as assessed by the incoming nurse over the course of the next shift). In this case, a time delayed measure of trust may reveal the hypothesized link between non-linguistic elements of speech and trust.

Our study contributes to trust research by suggesting that trust influences effective communication through a non-linguistic path. Although substantial research on trust suggests that trust facilitates communication and information sharing (Currall and Judge, 1995), there is little if any work suggesting that trust improves communication by facilitating non-linguistic elements of speech that, in turn, enhance information transfer.

---

[2] Joint Commission on Accreditation of Healthcare Organizations. Root causes of sentinel events. 2004. http://www.jcaho.org/accredited+organizations/ambulatory+care/sentinel+events/root+causes+of+sentinel+event.htm

We did not find a significant correlation between vocal features and trust during the observed transition in care, although we did find a significant correlation between vocal features and trust measured several weeks earlier. We therefore believe that the significance of vocal features for trust is still untapped. For instance, vocal features may only be important for trust in new relationships or after a trust violation. In these contexts, such features may signal genuine interest and engagement in the relationship. Because of their potential signaling value, vocal features may play an important role in trust repair. For instance, they may moderate the impact of trust repair strategies such as apologies and accounts on subsequent levels of trust.

The benefits of the social signaling methodology for predicting persuasion, interest, and handoff success are compelling. Wider application of this computer technique to trust research is demanded not only due to this success but also because of the relative ease with which these features can be extracted, especially when compared to manual coding. In the future we hope that additional sensors, such as accelerometers, infra-red transceivers and the like will be used by researchers to develop even richer datasets. Armed with these new analytical tools, we are sure that future research will yield many unprecedented and useful results.

References

Ambady, N., LaPlante, D., Nguyen, T., Rosenthal, R., Chaumenton, N., & Levinson, W. (2002), 'Surgeons' tone of voice: A clue to malpractice history', *Surgery* , pp. 5-9.

Bailenseon, J., & Yee, N. (2005), 'Digital Chameleons: Automatic Assimilation of Nonverbal Gestures in Immersive Virtual Environments', *Psychological Science* , pp. 814-819.

Bailenson, J. N., Iyengar, S., & Yee, N. (2005), 'Facial Identity Capture and Presidential Candidate Preference', *55th Annual Conference of the International Communication Association.*

Basu, S. (2002), *Conversational Scene Analysis*, MIT Media Lab PhD Thesis (Advisor Prof. Alex Pentland).

Chartrand, T. L., & Bargh, J. A. (1999), 'The Chameleon Effect: The Perception-Behavior Link and Social Interaction', *Journal of Personality and Social Psychology* , pp. 893-910.

Choudhury, T. (2004), *Sensing and Modeling Human Networks*, Cambridge, MA USA: PhD Thesis, MIT Media Laboratory.

Curhan, J., & Pentland, A. (2007), 'Thin Slices of Negotiation: Predicting Outcomes From Conversational Dynamics Within the First 5 Minutes', *Journal of Applied Psychology* , pp. 802-811.

Currall, S. C., & Judge, T. A. (1995), 'Measuring trust between organizational boundary role persons', *Organizational Behavior and Human Decision Processes*, vol. 64, pp. 151-170.

Edmondson, A. (1999), 'Psychological safety and learning behavior in work teams', *Administrative Science Quarterly*, **44** (2), 350-383.

Jaffee, J., Beebe, B., Feldstein, S., Crown, C. L., & Jasnow, M. (2001), 'Rhythms of dialogue in early infancy', *Monographs of the Society for Research in Child Development*.

Levin, D.Z. and R. Cross. (2004), 'The strength of weak ties you can trust: The mediating role of trust in effective knowledge transfer', *Management Science*, **50** (11), 1477-1490.

Madan, A., Caneel, R., & Pentland, A. (2004), *Voices of Attraction.* MIT Media Laboratory Technical Report.

Mayer, R.C. and J.H. Davis. (1999), 'The effect of the performance appraisal system on trust for management: A field quasi-experiment', *Journal of Applied Psychology*, **84**, 123-136.

Olguin Olguin, D., Waber, B., Kim, T., Mohan, A., Ara, K., & Pentland, A. (2009), 'Sensible Organizations: Technology and Methodology for Automatically Measuring Organizational Behavior', *IEEE Transactions on Systems, Man, and Cybernetics Part B* , pp. 43-55.

Pentland, A. (2008), *Honest Signals.* The MIT Press.

Pentland, A. (2004), *Social Dynamics: Signals and Behavior*, ICDL. IEEE Press.

Rabiner, L. R. (1989), 'A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition', *Proceedings of the IEEE* , pp. 257-286.

Stoltzman, W. T. (2006), *Toward a Social Signaling Framework: Activity and Emphasis in Speech,* Cambridge, MA, USA: MIT Media Laboratory Master's Thesis.

Wu, L., Waber, B. N., Aral, S., Brynjolfsson, E., & Pentland, A. (2008), *Mining Face-to-Face*

Interaction Networks using Sociometric Badges: Predicting Productivity in an IT Configuration

Task, ICIS 2008: Paris, France.

Suggested further reading

Basu, S. (2002), *Conversational Scene Analysis.* MIT Media Lab PhD Thesis (Advisor Prof. Alex Pentland).

Curhan, J., & Pentland, A. (2007), 'Thin Slices of Negotiation: Predicting Outcomes From Conversational Dynamics Within the First 5 Minutes', *Journal of Applied Psychology* , pp. 802-811.

Pentland, A. (2008), *Honest Signals*, The MIT Press.

Appendix A

Measures of Non-linguistic Vocal Signaling	
Activity Level	Calculation of the activity measure begins by using a two-level Hidden Markov Model (HMM)[3] to segment the speech stream of each person into voiced and non-voiced segments, and then group the voiced segments into speaking vs. non-speaking (Basu, 2002). Conversational activity level is measured by the z-scored percentage of speaking time plus the frequency of voiced segments.
Engagement	Engagement is measured by the z-scored influence each person has on the other's turn taking. Intuitively, when someone is trying to drive the conversation, they are more engaged than their conversational partner. When two people are interacting, their individual turn-taking dynamics influence each other and can be modeled as a Markov process (Jaffee, Beebe, Feldstein, Crown, & Jasnow, 2001). By watching people interact over long periods of time, we can determine what their normal interaction patterns are and see how they are influenced by the person they are currently interacting with. To measure these influences we model their individual turn-taking by an HMM and measure the coupling of these two dynamic systems to estimate the influence each has on the others' turn-taking dynamics (Choudhury, 2004). Our method is similar to the classic method of Jaffe et al. (Jaffee, Beebe, Feldstein, Crown, & Jasnow, 2001), but with a simpler parameterization that permits the direction of influence to be calculated and permits analysis of conversations involving many participants.
Emphasis	Emphasis is measured by the variation in prosodic emphasis. For each voiced segment we extract the mean energy, frequency of the fundamental format, and the spectral entropy. Averaging over longer time periods provides estimates of the mean-scaled standard deviation of the energy, formant frequency and spectral entropy. The z-scored sum of these standard deviations is taken as a measure of speaker stress; such stress can be either purposeful (e.g., prosodic emphasis) or unintentional (e.g., physiological stress caused by discomfort).

[3] An HMM is a statistical model that consists of a series of states, each of which is only dependent upon the previous state. Each state has a certain probability of outputting different symbols. After the model parameters have been chosen (number of states, possible state transitions, possible outputs), the Baum-Welch algorithm is used on a training set of data to find the optimal values of state transition and output probabilities based on the initial starting conditions of the model, which consists of initial state transition and output probabilities. A more detailed description of HMMs can be found in (Rabiner, 1989).

Appendix A (cont.)

Measures of Non-linguistic Vocal Signaling	
Mirroring	Mirroring behavior, in which the prosody of one participant is "mirrored" by the other, is considered to signal empathy, and has been shown to positively influence the outcome of a negotiation (Chartrand & Bargh, 1999). It has even been manipulated in the past in virtual reality experiments to influence the trustworthiness of avatars (Bailenseon & Yee, 2005). While we cannot measure mirroring directly using automated methods, we can look for mirroring like behavior by detecting short interjections ("uh-huh") and a quick exchange of words ("OK?", "OK!"). The z-scored frequency of these short utterance exchanges is taken as a measure of mirroring. In our data these short utterance exchanges were also periods of tension release.[4]

[4] When extracting time-dependent features such as mirroring and interruptions, time synchronization of the recorded data is essential. While under certain circumstances this may be done automatically, the most accurate method still relies on human intervention, introducing somewhat of a time burden on the researchers, although this requires much less time than manual coding.
Using combined sensor packages helps alleviate some of these issues. The recent development of Sociometric Badges pairs a microphone with a radio, clock, and other sensors to allow for the automatic synchronization of data by using the actual time transmitted by base stations when logging data (Olguin Olguin, Waber, Kim, Mohan, Ara, & Pentland, 2009).
Another method is to use easily recognizable, unique sounds on the audio track to aid automatic synchronization. Loud hand claps are particularly useful, since they are easy to generate and leave a distinct frequency signature in the audio data.

Author Bio

Benjamin Waber is a doctoral candidate in the Human Dynamics Group at the MIT Media Lab and a graduate student affiliate at the Institute for Quantitative Social Science at Harvard University. Benjamin has been at the forefront of applying sensor and electronic data to management research, studying problems such as informal communication structure, office layout, team behavior, and employee training.

His current research interests include dynamic organizational design, organizational behavior, social networks, sensor networks, prediction mechanisms, and information flow.

Michele Williams is an Assistant Professor of Organizational Behavior at Cornell University. Her research focuses on the development of cooperative, high-performance interpersonal relationships, especially on projects involving people from multiple organizations or groups within an organization. Williams' research concentrates on the influences of interpersonal processes, such as perspective taking, on how interpersonal trust and cooperation evolve.

Professor Williams has consulted on effective collaboration for public and private organizations such as Booz•Allen & Hamilton and Massachusetts General Hospital. She is also co-author of the 4-CAP Leadership Assessment—a 360° assessment used by organizations to enhance the leadership potential of managers.

John S. Carroll is a Professor of Behavioral and Policy Sciences at M.I.T.'s Sloan School of Management. Dr. Carroll conducts research on social-psychological and organizational factors that promote of safety in high-hazard industries such as nuclear power and health care. He focuses on: (1) safety culture as supported by communication, leadership, and systems thinking. (2) self-analysis and organizational learning. Dr. Carroll is a Fellow of the American Psychological Society and consultant to the Center of Excellence grant to the Harvard hospitals on quality of medical care and patient safety. He has published four books and numerous articles.

Professor Alex (Sandy) Pentland is a pioneer in organizational engineering, mobile information systems, and computational social science. Sandy's focus is the development of human-centered technology, and the creation of ventures that take this technology into the real world.

He directs the Human Dynamics Lab, helping companies to become more productive and creative through organizational engineering, and the Media Lab Entrepreneurship Program, which helps translate cutting-edge technology into real-world impact around the world. He is among the most-cited computer scientists in the world.