Real-time comprehensive sociometrics for two-person dialogs

Dauwels, Shoko; Rasheed, Umer; Tahir, Yasir; Dauwels, Justin; Thalmann, Daniel

2013

Magnenat-Thalmann, N., Rasheed, U., Tahir, Y., Dauwels, S., Dauwels, J., & Thalmann, D. (2013). Real-time comprehensive sociometrics for two-person dialogs. Proceedings of the 4th International Workshop, Human Behavior Understanding (HBU 2013), LNCS 8212, 196-208.

https://hdl.handle.net/10356/101166

https://doi.org/10.1007/978-3-319-02714-2_17

Real-Time Comprehensive Sociometrics for Two-Person Dialogs

Umer Rasheed[*], Yasir Tahir[‡], Shoko Dauwels[†], Justin Dauwels[*], Daniel Thalmann[‡], Nadia Magnenat-Thalmann[‡]

[*]School of Electrical and Electronic Engineering (EEE),
[†]CIRCQL, Nanyang Business School, [‡]Institute for Media Innovation (IMI), Nanyang Technological University, Singapore

Abstract. A real-time system is proposed to quantitatively assess speaking mannerisms and social behavior from audio recordings of two-person dialogs. Speaking mannerisms are quantitatively assessed by low-level speech metrics such as volume, rate, and pitch of speech. The social behavior is quantified by sociometrics including level of interest, agreement, and dominance. Such quantitative measures can be used to provide real-time feedback to the speakers, for instance, to alarm to speaker when the voice is too strong (speaking mannerism), or when the conversation is not proceeding well due to disagreements or numerous interruptions (social behavior). In the proposed approach, machine learning algorithms are designed to compute the sociometrics (level of interest, agreement, and dominance) in real-time from combinations of low-level speech metrics. To this end, a corpus of 150 brief two-person dialogs in English was collected. Several experts assessed the sociometrics for each of those dialogs. Next, the resulting annotated dialogs are used to train the machine learning algorithms in a supervised manner. Through this training procedure, the algorithms learn how the sociometrics depend on the low-level speech metrics, and consequently, are able to compute the sociometrics from speech recordings in an automated fashion, without further help of experts. Numerical tests through leave-one-out cross-validation indicate that the accuracy of the algorithms for inferring the sociometrics is in the range of 80-90%. In future, those reliable predictions can be the key to real-time sociofeedback, where speakers will be provided feedback in real-time about their behavior in an ongoing discussion. Such technology may be helpful in many contexts, for instance in group meetings, counseling, or executive training.

1 Introduction

In recent years, human behavior has gained much attention in the information sciences community. Specifically, automatic analysis of human behavior has become a major research topic, because of its important potential applications and its scientific challenges. Human behavior involves various patterns of actions and activities, attitudes, affective states, social signals, semantic descriptions, and contextual properties [1]. A promising approach to human behavior understanding is to apply pattern recognition and modeling techniques to automatically and objectively deduce various aspects human behavior from different kinds of recordings and measurements, e.g., audio and video recordings [2].

Speech analysis is one of the most common ways to analyze human behavior. Speaking mannerisms are a direct manifestation of human behavior, and play a vital role for the meetings to be pleasant, productive, and efficient [3]. If speaking mannerisms become mutually compatible and aligned, the meetings are more likely to be productive [4]. Appropriate feedback on speaking mannerisms and behavior during conversations may help to improve communications skills. This is especially true for real-time feedback, which allows the speaker to adjust on the fly.

In this paper, we aim to develop real-time algorithms to automatically quantify a variety of speaking mannerisms and social behavior in two-person dialogs for real-time sociofeedback (see Fig.1). Quantitative measures can provide a comprehensive picture of the behavior of participants in dialogs. In the following, we will briefly review related studies, and will highlight the novelties and contributions of this paper.

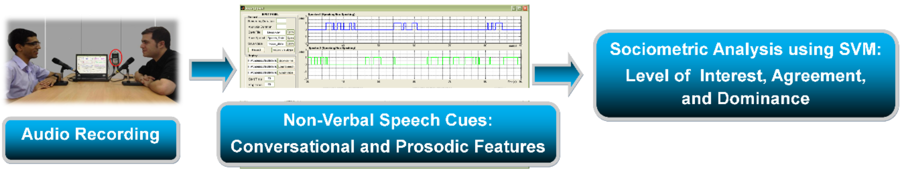
Audio Recording
Non-Verbal Speech Cues:
Conversational and Prosodic Features
Sociometric Analysis using SVM:
Level of Interest, Agreement,
and Dominance

Fig. 1. System Overview. The system records audio data, computes several conversational and prosodic features, and from those features, computes levels of interest, agreement and dominance via SVM.

Several recent studies consider modeling and automatic detection of personality traits, social relations and social roles etc from speech recordings [5-14]. In [12-13], several efforts have been made in automatic detection of speaker traits, social signals, conflict, emotions etc from short clips. Also automatic detection of interest or activity levels in multi-party dialogs has recently been explored [15-16]. The concept of hot-spots is introduced in [16] to identify periods of high interest in meeting recordings. In [17], emphasis and interest in conversations is inferred from speech pitch. Furthermore, in [18] a PDA-based system is proposed to extract interest levels in conversations. Interest level has also been investigated in computer-assisted learning [19]. Similarly, detection of agreement has been investigated in meeting scenarios such as broadcast conversations [20-23]. In most of those studies, the proposed algorithms are evaluated on the annotations from the ICSI corpus [22]. Dominance and similar concepts like emerging leader have been investigated in a similar manner. In social psychology, dominance is usually interpreted either as a personality characteristic or in relation to the hierarchical position of an individual within a group [24]. In [25-26], methods are proposed that detect dominance in multi-party dialogs in an automated fashion from non-verbal audio and visual cues.

The current studies on automated behavioral analysis of dialogs have the following limitations:

- In most studies, only one aspect of social behavior in dialogs is considered, e.g., dominance or activity level [15-25].
- Most of the recent literature only considers two levels of social behavior, e.g., active vs. non-active, or dominant vs. non-dominant [15, 19, 25, 26]. However, there have been studies that perform multi-level classification of personality traits, interest, conflict etc from short clips of 10-30s using lexical features [11-13].

- Typically, the speech corpus used for analysis and testing purposes is limited in scope, and mostly contains a specific kind of dialog, e.g. broadcast shows [9-11, 15-23, 26]. In most dialogs analyzed so far, the speakers tend to cooperate and the conversations proceed smoothly. Problematic scenarios such as conflicts, disagreements and boredom have received little attention [12].

In this paper, we introduce the following contributions and novelties:

- We consider multiple aspects of behavior in dialogs simultaneously, including a variety of speaking mannerisms (e.g., volume, pitch, and rate of speech) and social behavior (interest, agreement, dominance).
- Instead of binary classification of behavior (e.g., dominant vs. non-dominant), we consider multi-level classification (low, slightly low, normal, slightly high, high). As a result, the proposed system provides more refined feedback.
- We have collected a novel annotated speech corpus, in English, that spans a wide range of brief dialogs, including also problematic situations such as conflicts and disagreements, periods of boredom, aggressive behavior, or poorly delivered speech (e.g., low voice or fast pace). Each dialog in the corpus has been assessed by at least 5 people with regard to speaking mannerisms and social behavior (interest, agreement, dominance).
- The proposed automated algorithms have low computational complexity and can readily be used in real-time. Therefore, they can provide feedback while the dialog is still ongoing. Sociofeedback can be provided via many platforms, for example, smartphones, voice over IP (e.g., Skype), or humanoid robots (e.g., Nao robot).
- Our objective in the long term is to provide real-time feedback on social behavior in dialogs, for training purposes and also for therapy of psychiatric patients with deficits in social behavior.

Specifically, we propose the following approach. First, low-level speech cues are extracted from the audio recordings, e.g., volume, rate, and pitch of speech. We apply feature selection methods such as Information Gain (IG) and Correlation-based Feature Selection (CFS) to determine the most relevant low-level speech cues for quantifying social behavior [27-28]. We then train machine learning algorithms such as support vector machines (SVM) [29] with those features as inputs, to quantify certain aspects behaviors, i.e., level of interest, agreement, and dominance. We collected and annotated a new speech corpus to train the machine learning algorithms. Through this training procedure, the algorithms learn how the three sociometrics (level of interest, agreement, and dominance) depend on the low-level speech metrics, and consequently, are able to compute the sociometrics from speech recordings in an automated fashion in real-time, without further help of experts. The real-time sociometrics can be helpful to provide feedback to participants in dialogs.

In this paper, we limit ourselves to behavior detection of face-to-face two-person conversations. In the future, we intend to scale our system towards small-group interactions. We also plan to incorporate visual cues alongside, in order to attain more detailed sociometrics. The ultimate goal is not limited to feature selection/pattern recognition/classicification but development of a system that can provide sociofeedback for different types of social interactions, e.g. job interviews, business meetings, or group discussions [30].

The paper is structured as follows. In Section 2, we describe the collected speech corpus and our experimental setup. In Section 3, we elaborate on the low-level speech metrics. In Section 4, we specify the sociometrics, and explain our machine learning approach to automated prediction of sociometrics. In Section 5 we offer concluding remarks, and suggest several topics for future research.

2 Speech corpus

The newly collected speech corpus contains 150 two-person conversations, each at least 2.5-3 minutes long. Consequently, the dataset consists of 300 individual audio recordings in total. The total number of individuals participating in the corpus is 22, of which 17 are males and 5 are females. The participants are students of Nanyang Technological University (NTU). The age of the students is from 18 to 30 (M=25, SD=1.20). The topics of conversations ranged from discussion of assignments, projects of students, to social and political views. In some of the dialogs, there are problematic situations such as conflicts and disagreements, periods of boredom, aggressive behavior, or poorly delivered speech (e.g., low voice or fast pace). The length of each recording is relatively long (2-3 minutes) as compared to other corpora [13-14, 22]. The 2-3 minute conversations provide overall sense of sociometrics and hence are useful in studying the design of appropriate sociofeedback. In the following, we discuss how we recorded the corpus, and how it has been annotated.

2.1 Experimental Procedure

The following section explains the procedure of experiment in detail:

1. First, we setup the recording system properly. We adopt easy-to-use portable equipment for recording conversations; it consists of lapel microphones for each of the two speakers and an audio H4N recorder that allows multiple microphones to be interfaced with the computer. The audio data is recorded in brief consecutive segments and sent to a central server running MATLAB. The speech is saved in a 2-channel audio .wav file in order to allow precise computation of overlap-related features such as interruptions.
2. The two participants sit about 1.5m apart so that each microphone only records the voice of the respective participant, and there is no interference from the other participant.
3. We attach the microphones to the participants in proper manner, in order to obtain a high-quality recording.
4. We brief the participants about the experiment. We provide them a list of topics that they can choose from. The two participants are asked to agree on a topic for the subsequent discussion. The topics of discussion range from small talk to heated debates on sports, politics, etc. We selected the topics of discussion carefully in order to evoke a variety of behaviors.
5. Only the two participants remain in the meeting room. The recording system is controlled remotely through a wireless connection.
6. We start the recording via a laptop remotely connected to the server.

7. We monitor the conversation remotely for about 2.5-3 minutes at least, without any interruption. Participants are allowed to talk as long as they wish. As the participants do not need to keep track of time, they can freely engage in the conversation.

2.2 Manual Annotation

Each audio recording in the corpus is annotated by a pool of at least 5 people ("judges"). There were 14 judges in total, each assessing a subset of the corpus. For each audio recording, the judges completed a questionnaire (see Table 1) related to speaking mannerisms and behavioral aspects of each participant. The responses range from 1 (low) to 5 (high). For example, if a participant seems bored, his interest level is annotated as "low"; in contrast, if the participant seems excited, then the interest level is assessed as "high". Table 2 shows the standard deviation of the annotations for the different values of each metric. As can be seen from that table, the standard deviation remains low for most of the classes, suggesting there is no large disparity among the judges. However, for certain classes the standard deviation is a little high.

Assessment of Speech Mannerism
This person was too loud.
This person was not audible.
This person screamed at other participants.
This person spoke too fast.
This person spoke too slow.
This person stuttered while speaking.
The speech clarity was satisfactory.
This person spoke in a monotonic way.
This person is not responsive.
This person interrupted the other participant(s).
Assessment of Social Behavior
This person seemed to be actively engaged in the conversation.
This person seemed to agree with what other person has to say.
This person seemed to be the dominant of the two.

Table 1. Questionnaire for sociometric assessment.

Category	Low	Slightly Low	Normal	Slightly High	High
Interest	0.71	0.96	1.10	1.15	0.63
Agreement	0.63	0.84	0.74	0.92	0.64
Dominance	0.54	1.03	0.51	0.67	0.57

Table 2. Standard deviation of manual annotation by multiple judges, for each level of interest, agreement, and dominance. "Low" corresponds to an average score below 1.5, "slighly low" is associated with an average score between 1.5 and 2.5, and so on.

3 Inferring Speech Mannerisms from Speech Cues

We consider two types of low-level speech metrics: conversational and prosody related cues. The conversational cues account for *who* is speaking, *when* and *how much*, while the prosodic cues quantify *how* people talk during their conversations. Non-verbal speech cues play a significant role in group conversations [9-11]. In the following, we briefly review the conversational and prosodic cues that we consider in this study. Then we explain how we infer speaking mannerisms from those cues.

3.1 Conversational Cues

In order to compute the conversational features, we first perform speech detection by means of a hidden Markov model (HMM) that uses energy-independent features [31]. This approach to speech detection is robust to low sampling rates, significant levels of noise, and variations in the distance between speaker and microphone. The speech detection algorithm works in real-time. The time taken for 2-3 min dialog takes 5-8 seconds on a laptop with 2GHz dual-core processor and 2GB RAM. Once the audio signals have been segmented in periods of speech and without speech, we compute the following conversational cues: the number of natural turns, speaking percentage, mutual silence percentage, turn duration, interjections, interruptions, failed interruptions, and response time (see Fig.2).

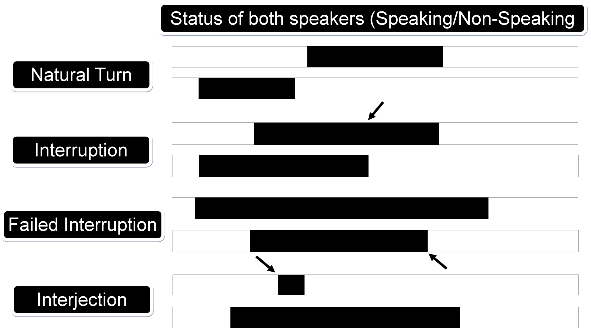

Fig. 2. Illustration of turn-taking, interruption, failed interruption, and interjection. Those conversational cues are derived from the binary speaking status (speaking vs. non-speaking). Periods of speaking and non-speaking are indicated in black and white respectively [30].

3.2 Prosodic Cues

We consider the following prosodic cues: amplitude, larynx frequency (F0), formants (F1, F2, F3), and mel-frequency cepstral coefficients (MFCCs); those cues are extracted from 30ms segments at a fixed interval of 10ms. Those cues fluctuate rapidly in time. Therefore, we compute various statistics of those cues over a time period of several seconds, including minimum, maximum, mean and entropy, in order to assess speaking mannerisms.

3.3 Speech Mannerism

From the speech cues, we assess a variety of speech mannerisms in real-time. This quantitative information can be provided to the speaker as feedback about an ongoing dialog. Specifically, we aim to detect the following speech mannerisms: the speech volume is too low/high, the speaker is screaming, the speech rate is too low/high, and the speaker is taking too much time to respond. The system monitors relevant speech cues for each speech mannerism (see Table 3), and checks whether any cue is abnormally low or high. For instance, low (high) speech volume is detected when the speech signal volume is below (above) a certain threshold. Likewise, fast speaking is detected when the speech rates is above a certain threshold. In order to set those thresholds, we analyze short clips of speech of approximately 10s each. From the scores provided by the judges, we can define speech mannerisms. For instance, when the average score for "This person was too loud" is 4.5 or higher, the speaker is considered to be speaking too loud. Similarly, when the average score for "This person spoke too slow" is 4.5 or higher, the speaker is considered to be speaking too slow.

Speech Mannerism	**Corresponding Feature(s)**	**Detection Rate(%)**
Low Voice	Mean Volume	88
High Voice	Mean Volume	90
Low Speech Rate	Speech Rate	76
High Speech Rate	Speech Rate	78
Screaming	Mean Volume, Mean MFCCs	92
Long Response Time	Response Time	96

Table 3. Prediction level for speech mannerisms.

As we mentioned earlier, in order to detect such mannerisms from the speech cues, we choose certain thresholds. For example, according to the annotations of the speech corpus, a volume level below 30dB is perceived as too silent, and a volume level of above 80dB is usually considered too loud. We select the values of the thresholds such that the false-alarm rate is about 5%.

Speech Mannerism	Low	Normal	High	Classified as
Volume	22	2	0	Low
	3	36	2	Normal
	0	2	18	High
Speech Rate	19	1	0	Low
	6	23	3	Normal
	0	1	17	High

Table 4. Confusion matrix for speech mannerisms: speech volume and rate.

It is noteworthy that the speech cues can be computed in real-time. That also holds for detecting speech mannerism from those cues. Therefore, if any of these speech cues fall above or below a certain threshold, then the system can provide feedback in real-time to the participant. For instance, if the speech volume is too high, the system can inform the participant to lower the voice and vice versa.

Speech Mannerism	True	False	Classified as
Screaming	23	2	True
	2	48	False
Long Response Time	19	2	True
	1	48	False

Table 5. Confusion matrix for speech mannerisms: screaming and long response time.

Our results are summarized in Table 3-5. In Table 3, we list the sensitivity (detection rate) of the threshold-based detector for different speech mannerisms. Tables 4-5 show the confusion matrices for the classification of speech mannerisms. Overall, the results suggest that basic speech mannerisms can be detected accurately from speech cues.

4 Sociometrics

In this section, we describe how we quantify social behavior in dialogs. Specifically, we compute three sociometrics: level of interest, agreement, and dominance. Those sociometrics are inferred from combinations of speech cues. First we discuss how we can assess the relevance of the different speech cues for the three sociometrics. Next we explain how we infer the sociometrics in an automated fashion from selected speech cues.

4.1 Feature Selection

The speech cues tend to be highly correlated with some sociometrics, and totally uncorrelated with others. Consequently, it is crucial to select appropriate speech cues in order to predict the sociometrics. We apply two feature selection algorithms, viz. Information Gain (IG) and Correlation-based Feature Selection (CFS) [27-28], to determine the most relevant features for inferring each of the three sociometrics. In the following, we briefly review these two methods.

Information gain is one of the simplest of the attribute ranking algorithms. If A is a particular feature and C is the class (one of the 5 values of the sociometric), then the entropy $H(C)$ and $H(C|A)$ of the class before and after observing the feature respectively can be written as:

$$H(C) = -\sum_{c\in C} p(c)\log_2(p(c)), \tag{1}$$

$$H(C|A) = -\sum_{a\in A} p(a)\sum_{c\in C} p(c)\log_2(p(c)). \tag{2}$$

The amount by which the entropy of a particular class decreases is a measure of to the amount of information about the class provided by the feature, and is referred as information gain. Hence each feature A_i can be scored by means of its information gain [23]:

$$IG_i = H(C) - H(C|A_i) = H(A_i) - H(A_i|C) = H(A_i) + H(C) - H(A_i,C). \tag{3}$$

Correlation feature selection (CFS), on the other hand, evaluates subsets of features rather than individual features. This method accounts for the usefulness of individual

Category	Feature	Info Gain	CFS Subset Merit
Interest	Speaking %	1.130	0.718
	Turn Duration	0.564	0.540
	Mutual Silence	0.468	0.392
	Volume	0.420	0.312
	Response Time	0.392	0.192
Agreement	Interruptions	0.888	0.605
	Total Overlap	0.508	0.400
	Mutual Silence	0.227	0.191
	Larynx Frequency (F0)	0.212	0.166
	Volume	0.167	0.218
Dominance	Speaking % Difference	1.079	0.783
	Turns Difference	0.695	0.540

Table 6. Information gain and CFS subset merit of the features used for optimal classification.

features for predicting the class along with the level of inter-correlation between them. The CFS subset merit is high when the features in the subset are highly correlated with the class, and have low inter-correlation with each other. CFS subset merit is calculated as:

$$\text{Merit}_S = \frac{k\bar{r}_{cf}}{\sqrt{k+k(k-1)\bar{r}_{ff}}}, \tag{4}$$

where Merit_S is the heuristic merit of a feature subset S containing k features, $\bar{r}_{cf}$ is the average feature-class correlation whereas $\bar{r}_{ff}$ is the average feature-feature inter-correlation [27].

Table 6 shows the results the Information Gain and CFS subset Merit for feature sets that yield best classification results. Usually the feature-sets include many features [32]. However, in such approach it becomes unclear what the key factors are, as the classifiers have many features as input. In this work, the feature-set for classification is kept small so that in future the system could be implemented for real-time sociofeedback on smartphones and humanoid robots etc. The table indicates that speaking percentage, turn duration, mutual silence percentage, volume, and response time are useful features for inferring the interest level; difference of turns and speaking percentage are good features for quantifying dominance; whereas interruptions, overlap, mutual silence percentage, volume, and F0 are relevant for quantifying agreement.

4.2 Multi-Class Classification

The sociometrics, including level of interest, agreement, and dominance, can take five values (1, 2, . . . , 5). That value is estimated from relevant speech cues (cf. Table 6). Specifically, we view this problem as multi-class classification, where the number of class equal five.

We train multi-class classifiers in a supervised manner using the newly collected speech corpus. The (rounded) average score provided by the judges serve as label for

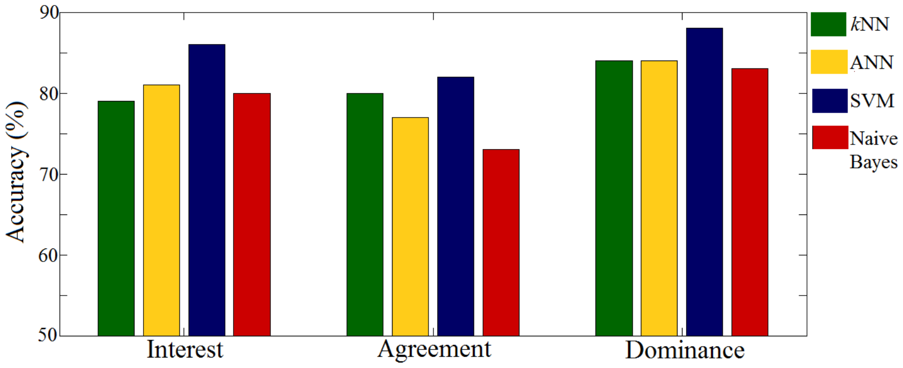

Fig. 3. 5-class classification performance for kNN, ANN, SVM, and naive Bayes.

supervised learning. We consider four kinds of multi-class classifiers: *k*-nearest neighbor, Artificial Neural Network (ANN), Naive Bayes, and SVM [29, 33-35]. The 5-class classification performance of these four algorithms for each sociometric is summarized in Fig.3. The classification performance is computed by leave-one-out crossvalidation i.e. each sample *n* is classified by all of the instances in the training set *N* other than *n* itself, so that almost all of the data is available for each classification attempt. As can be seen from Fig.3, the SVM performs better than the three other classifiers for all three sociometrics. The confusion matrices obtained by SVM for level of interest, agreement and dominance are displayed in Table 7. As can be seen from Fig.3 and Table 7, the values of the sociometrics can be inferred reliable from the speech cues by SVM.

Now we will briefly address the computational complexity of our approach. It takes approximately 5 seconds to train each SVM, 5-10 seconds to compute the speech detection and speech cues from 2-3 min dialogs, and less than a second to perform multi-class classification by SVM. In order to infer sociometrics, the speech cues need to be calculated, followed by classification. The total time required for inferring the sociometrics from a 2-3 min dialog is therefore about 5-10 seconds using 2GB dual-core processor and 2GB RAM. Consequently, the sociometrics can be inferred in real-time, and as a result, they maybe used to provide feedback to the speakers in real-time, while the dialog is ongoing.

5 Discussion and Conclusion

In this paper, we presented a novel approach towards comprehensive real-time analysis of speech mannerism and social behavior. Using low-level speech metrics, we quantified speech mannerisms and sociometrics, i.e., interest, agreement and dominance of the speaker. We collected a diverse speech corpus of two-person conversations; it allowed us to train machine learning algorithms for reliable 5-level classification of the sociometrics with speech cues as input features.

The combined metrics for speech mannerisms and social behavior provide a clear picture of human behavior in dialogs. For instance, high volume, high dominance coupled with low agreement level may suggest that the participant is upset and behaving

Behavior	Low	Slightly Low	Normal	Slightly High	High	Classified as
Interest	32 (91%)	7 (9%)	0 (0%)	0 (0%)	0 (0%)	Low
	3 (9%)	60 (80%)	7 (7%)	0 (0%)	0 (0%)	Slightly Low
	0 (0%)	8 (11%)	88(87%)	6 (13%)	0 (0%)	Normal
	0 (0%)	0 (0%)	6 (6%)	36 (80%)	6 (14%)	Slightly High
	0 (0%)	0 (0%)	0 (0%)	3 (7%)	38 (86%)	High
	0.27	0.45	0.36	0.45	0.37	RMSE
Agreement	22 (78%)	3 (5%)	0 (0%)	0 (0%)	0 (0%)	Low
	5 (18%)	49 (82%)	10 (11%)	0 (0%)	0 (0%)	Slightly Low
	1 (4%)	8 (13%)	74 (79%)	4 (7%)	2 (3%)	Normal
	0 (0%)	0 (0%)	9 (10%)	45 (79%)	5 (8%)	Slightly High
	0 (0%)	0 (0%)	0 (0%)	8 (14%)	55 (89%)	High
	0.38	0.46	0.45	0.43	0.50	RMSE
Dominance	45 (93%)	3 (4%)	0 (0%)	0 (0%)	0 (0%)	Low
	3 (7%)	67 (84%)	4 (5%)	0 (0%)	0 (0%)	Slightly Low
	0 (0%)	10 (12%)	74 (87%)	6 (13%)	0 (0%)	Normal
	0 (0%)	0 (0%)	7 (8%)	39 (81%)	4 (10%)	Slightly High
	0 (0%)	0 (0%)	0 (0%)	3 (6%)	35 (90%)	High
	0.25	0.40	0.36	0.44	0.32	RMSE

Table 7. Confusion matrix for level of interest, agreement, and dominance.

aggressively. Similarly, low volume, low interest and low dominance may suggest that the participant is bored. More research is required to interpret the interplay between the three sociometrics.

Although preliminary, the results are encouraging: the sociometrics can be computed fast and reliably, enabling real-time sociofeedback. In current work, we are exploring applications of sociofeedback. In the future, we will include video recordings in conjunction with audio recordings. Our current speech corpus contains 150 conversations only. We intend to collect a much larger, diverse dataset, in order to generalize the findings. Also, the current work is limited to two-person face-to-face dialogs. In the future, we will scale the system to multi-party dialogs. The ultimate aim in this line of research is to develop template-based sociofeedback system for various types of social interactions, e.g., interviews, coaching sessions, group discussions.

Acknowledgements

This research project is supported in part by the Institute for Media Innovation (Seed Grant M4080824) Nanyang Technological University (NTU).

References

1. Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A. (eds.): Human Behavior Understanding, First International Workshop, HBU 2010, Istanbul, Turkey, August 22, 2010. Proceedings, Lecture Notes in Computer Science, vol. 6219. Springer (2010)
2. Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A.: Challenges of human behavior understanding. In: Human Behavior Understanding, First International Workshop, HBU 2010. Proceedings, Lecture Notes in Computer Science, vol. 6219. pp. 1-12. Springer (2010)

3. Pentland, A.S.: Honest Signals: How They Shape Our World, MIT Press (2008)
4. Pentland, A.S.: Socially aware, computation and communication, Computer, vol. 38, no. 3, pp. 3340 (2005)
5. Barzilay, R., Collins, M., Hirschberg, J., Whittaker, S.: The rules behind roles: Identifying speaker role in radio broadcasts. In: Proceedings of Association for the Advancement of Artificial Intelligence (AAAI), pp. 679684 (2000)
6. Liu, Y.: Initial study on automatic identification of speaker role in broadcast news speech. In: Proceedings of HLT/NAACL, pp. 8184 (2000)
7. Hutchinson, B., Zhang, B., Ostendorf, M.: Unsupervised broadcast conversation speaker role labeling. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 53225325 (2010)
8. Salah, A.A., Lepri, B., Pianesi, F., Pentland, A.: Human Behavior Understanding for Inducing Behavioral Change: Application Perspectives. In: Human Behavior Understanding, HBU 2011. Proceedings, Lecture Notes in Computer Science, vol. 7065, pp 1-15. Springer (2011)
9. Vinciarelli, A., Salamin, H., Polychroniou, A., Mohammadi, G., Origlia, A.: From Nonverbal Cues to Perception: Personality and Social Attractiveness. COST 2102 Training School pp. 60-72 (2011)
10. Pianesi, F., Zancanaro, M., Not, E., Leonardi, C., Falcon, V., Lepri, B.: Multimodal support to group dynamics. In: Proceedings of Personal and Ubiquitous Computing, vol. 12, no. 3, pp. 181-195 (2008)
11. Mohammadi, G., Mortillaro, M., Vinciarelli, A.: The Voice of Personality: Mapping Nonverbal Vocal Behavior into Trait Attributions. In: Proceedings of the International Workshop on Social Signal Processing, pp. 17-20, 2010
12. Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Sherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., Kim, S.: The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. INTERSPEECH (2013)
13. Schuller, B., Steidl, S., Batliner, A., Noth, E., Vinciarelli, A., Burkhardi, F., Son, R.V., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G., Weiss, B.: The INTERSPEECH 2012 Speaker Trait Challenge, INTERSPEECH (2012)
14. Nishimura, R., Kitaoka, N., Nakagawa, S.: Analysis of relationship between impression of human-to-human conversations and prosodic change and its modeling. INTERSPEECH, pp. 534-537, 2008
15. Gatica-Perez, D., McCowan, I., Zhang, D., and Bengio, S.: Detecting Group Interest-Level in Meetings. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 489-492 (2005)
16. Hornler, B., Rigoll, G.: Multi-modal activity and dominance detection in smart meeting rooms. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1777-1780 (2009)
17. Kennedy, L., Ellis, D.: Pitch-based emphasis detection for characterization of meeting recordings. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 243-248 (2003)
18. Eagle, N., Pentland, A.: Social network computing. In: Proceedings of UBICOMP, pp. 289-296 (2003)
19. Germesin, S., Wilson, T.: Agreement detection in multiparty conversation. In: Proceedings of the 2009 International Conference on Multimodal Interfaces (ICMI), pp. 7-14 (2009)
20. Wang, W., Yaman, S., Precoda, K., Richey, C.: Automatic identification of speaker role and agreement/disagreement in broadcast conversation. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5556-5559 (2011)

21. Kim, S., Valente, F., Vinciarelli, A.: Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5089-5092 (2012)
22. Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., Wooters, C.: The ICSI Meeting Corpus. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1. pp. 364-367 (2003)
23. Hillard, D., Ostendorf, M., and Shriberg, E.: Detection of Agreement vs. Disagreement in Meetings: Training with Unlabeled Data. In: Proceedings of HLT/NAACL, vol. 2. pp. 34-36 (2003)
24. Kalimeri, K., Lepri, B., Aran, O., Jayagopi, D.B., Gatica-Perez, D., Pianesi, F.: Modeling dominance effects on nonverbal behaviors using granger causality. In: Proceedings of the 2009 International Conference on Multimodal Interfaces (ICMI), pp. 23-26 (2012)
25. Rienks, R., Heylen, D.: Dominance Detection in Meetings Using Easily Obtainable Features In: Proceedings of Workshop on Machine Learning for Multimodal Interaction (MLMI), Lecture Notes in Computer Science, vol. 3869, pp. 76-86 (2006)
26. Wang, W., Precoda, K., Hadsell, R., Kira, Z., Richey, C., G. Jiva, G.: Detecting leadership and cohesion in spoken interactions. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 51055108 (2012)
27. Hall, M.A.: Correlation-based feature selection for machine learning. Ph.D. thesis, The University of Waikato, 1999.
28. Yu, L., Liu, H.: Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Proceedings of Machine Learning-International Workshop Then Conference, vol. 20, p.856 (2003)
29. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, vol. 2, no. 2, pp. 121167 (1998)
30. Sarda, S., Constable, M., Dauwels, J., Dauwels, S., Elgendi, M., Mengyu, Z., Rasheed, U., Tahir, Y., Thalmann, D., Magnenat-Thalmann, N.: Real-Time Feedback System for Monitoring and Facilitating Discussions, Natural Interaction with Robots, Knowbots and Smartphones - Putting Spoken Dialog Systems into Practice. Lecture Notes in Computer Science, Springer (2013)
31. Basu, S.: A linked-hmm model for robust voicing and speech detection. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1. pp. I-816-I-819 (2003)
32. Eyben, F., Wollmer, M., Schuller, B.: openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In Proceedings of ACM Multimedia (MM), pp. 1459-1462 (2010)
33. Wang, J., Zucker, J.D.: Solving Multiple-Instance Problem: A Lazy Learning Approach. In: Proceedings of 17th International Conference on Machine Learning (ICML), pp. 1119-1125 (2000)
34. Haykin, S.: Neural Network, A comprehensive foundation. Neural Networks, vol. 2 (2004)
35. Rish, I.: An empirical study of the naive Bayes classifier. In: Proceedings of IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, pp. 41-46 (2001)