

Voices of Attraction

Anmol Madan, Ron Caneel, Alex “Sandy” Pentland

MIT Media Laboratory
20 Ames Street, Cambridge 02139
{anmol, rcaneel, sandy}@media.mit.edu

Abstract

Non-linguistic social signals (e.g., ‘tone of voice’) are often as important as linguistic or affective content in predicting behavioral outcomes [1, 12]. This paper describes four automated measures of audio social signaling within the experimental context of speed dating. We find that this approach allows us to make surprisingly accurate predictions about the outcomes of social situations.

1 Introduction

In many situations non-linguistic social signals (body language, facial expression, tone of voice) are as important as linguistic content in predicting behavioural outcome [1, 12]. Tone of voice and prosodic style are among the most powerful of these social signals even though (and perhaps because) people are usually unaware of them [12]. In a wide range of situations (marriage counselling, student performance assessment, jury decisions, etc.) an expert observer can reliably quantify these social signals and with only a few minutes of observation predict about 1/3d of the variance in behavioural outcome (which corresponds to a 70% binary decision accuracy) [1]. In certain areas of human judgement these recorded predictions have been up to 90% accurate [7]. It is astounding that observation of social signals within such a ‘thin slice’ of behaviour can predict important behavioural outcomes (divorce, student grade, criminal conviction, etc.), even when the predicted outcome is sometimes months or years in the future.

Current literature in affective computing tries to link speaking style with specific emotional states. Picard [6,14], Oudeyer [13], and Brezal [3] look for observations of sympathetic or para-sympathetic nervous activity, which also cause variations in the intensity and pitch of voice (e.g. angry versus calm). This technique is well suited for applications that seek to understand the human’s internal state, for example, building computers that can measure frustration.

In contrast, we propose that minute-long averages of audio features often used to measure affect (e.g., variation in pitch, intensity, etc) taken together with conversational interaction features (turn-taking, interrupting, making sounds that indicate agreement like ‘uh-huh’) are more closely related to social signalling theory rather than to an individual’s affect. The ability to understand these social signals in an automated fashion allows us to make better predictions about the outcomes of *interactions* – speed-dating, negotiations, trading business cards --- than would affective measurements alone. Our reasoning is that social signals are actively constructed with communicative intent, whereas the connection between individual affect and group behaviour is extremely complex.

2 Measuring Social Signals

Pentland [15] constructed measures for four types of vocal social signalling, designated activity level, engagement, stress, and mirroring. These four measures were extrapolated from a broad reading of the voice analysis and social science literature, and we are now working to establish their general validity

Calculation of the activity measure begins by using a two-level HMM 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 [8,2]. Conversational activity level is measured by the z-scored percentage of speaking time plus the frequency of

voiced segments. The activity measure and stress measures are features common in affect recognition literature [6, 13].

Engagement is measured by the z-scored influence each person has on the other's turn taking. When two people are interacting, their individual turn-taking dynamics influence each other and can be modeled as a Markov process [9]. By quantifying the influence each participant has on the other we obtain a measure of their engagement...popularly speaking, were they driving the conversation? To measure these influences we model their individual turn-taking by a Hidden Markov Model (HMM) and measure the coupling of these two dynamic systems to estimate the influence each has on the others' turn-taking dynamics [5]. Our method is similar to the classic method of Jaffe et al. [9], but with a simpler parameterization that permits the direction of influence to be calculated and permits analysis of conversations involving many participants.

Stress 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 speaker stress; such stress can be either purposeful (e.g., prosodic emphasis) or unintentional (e.g., physiological stress caused by discomfort).

Mirroring behaviour, 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 [4]. In our experiments the distribution of utterance length is often bimodal. Sentences and sentence fragments typically occurred at several-second and longer time scales. At time scales less than one second there are short interjections (e.g., 'uh-huh'), but also back-and-forth exchanges typically consisting of single words (e.g., 'OK?', 'OK!', 'done?', 'yup.'). 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.

2.1 Signalling Dynamics

These measures of social signalling can be computed on a conventional PDA in real-time, using a one-minute lagging window during which the statistics are accumulated. It is therefore straightforward to investigate how these 'social signals' are distributed in conversation. In [10, 16] we analyzed social signalling in 54 hours of two-person negotiations on a minute-by-minute basis. We observed that high numerical values of any one measure typically occur by themselves, e.g., periods with high engagement do not show high stress, etc., so that each participant exhibits four 'social display' states, plus a 'neutral' relaxed state in which the participant is typically asking emotionally neutral questions or just listening. The fact that these display states were largely unmixed provides evidence that they are measuring separate social displays.

2.2 Attraction Experiment

Speed dating is relatively new way of meeting many potential matches during an evening. Participants interact for five minutes with their 'date', at the end of which they decide if they would like to provide contact information to him/her, and then they move onto the next person. A 'match' is found when both singles answer yes, and they are later provided with mutual contact information. Perhaps since speed-dating provides instant-gratification in the dating game, it has recently become the focus of several studies that attempt to predict outcomes based on mate selection theory [11]



Figure 1: Speed dating session in progress.
Audio is recorded using Zaurii PDAs.

In this experiment we analyzed 60 five-minute speed-dating sessions. Data was collected from several events held in 2004, and participants were singles from the MIT-Harvard community in the ages of 21-45. In addition to the 'romantically attracted' (provide contact information) question, participants were also asked two other yes/no questions: would they like to stay in touch just as friends, and would they like to stay in touch for a business relationship. These 'stay in touch' questions allowed us to explore whether romantic attraction could be differentiated from other factors.

Speed-dating is of particular experimental interest because there is a clear 'buying' decision at the end of each conversation. Since data was collected from real-life sessions, we were also more immune to the effect of self-entertainment. (i. e. we classify participants' natural intonation, not special 'interested' or 'bored' voices that actors generated).

2.3 Results

The four social signalling measures for both male and female were compared by linear regression to the question responses, and in each case the resulting predictor could account for more than 1/3rd of the variance. For the females responses, for instance, the correlation with the 'attracted' responses were $r=0.66$, $p=0.01$, for the 'friendship' responses $r=0.63$, $p=0.01$, and for the 'business' responses $r=0.7$, $p=0.01$. Corresponding values for the male responses were $r=0.59$, $r=0.62$, and $r=0.57$, each with $p=0.01$.

The engagement measure was the most important individual feature for predicting the 'friendship' and 'business' responses. The mirroring measure was also significantly correlated with female 'friendship' and 'business' ratings, but not with male ratings. The stress measure showed correlation with both participants saying 'yes' or both saying 'no' for the 'attraction' ($r=0.6$, $p=0.01$) and 'friendship' ($r=0.58$, $p=0.01$) questions.

An interesting observation was that for the 'attracted' question female features alone showed far more correlation with both male ($r=0.5$, $p=0.02$) and female ($r=0.48$, $p=0.03$) responses than male features (no significant correlation). In other words, female social signaling is more important in determining a couples 'attracted' response than male signaling. The most predictive individual feature was the female activity measure.

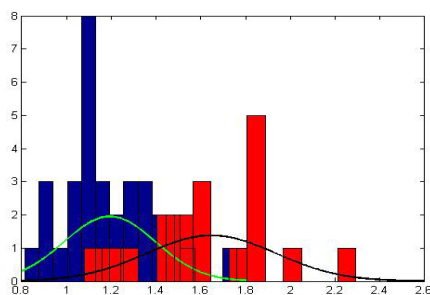


Figure 2: Frequency distribution of female 'attracted' responses (red=yes) vs. predictor value. The cross-validated binary linear decision rule has 72% accuracy

Figure 3 shows a two-class linear classifier for the 'attraction' responses, based on the social signaling measures; this classifier has a cross-validated accuracy of 71% for predicting the 'attracted' response. Feature selection was based on the regression results. The two fitted Gaussians are simply to aid visualization of the distributions' separability.

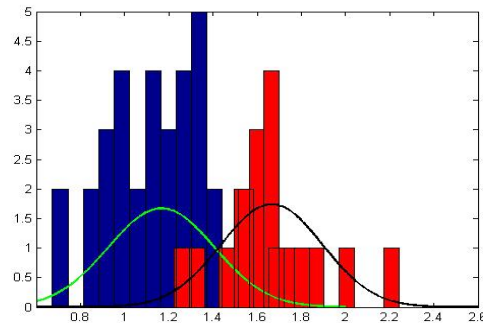


Figure 3: Frequency distribution of female 'business' responses (red=yes) vs. predictor value. The cross-validated three-class linear decision rule produces 83 % accuracy.

Figure 4 illustrates a two-class linear classifier for the 'business' responses, based on the social signaling measures; this classifier has a cross-validated accuracy of 74% for predicting the 'attracted' response. By considering the overlapping region as a third class, we can increase the cross-validation accuracy to 83% for the yes and no response regions. The two fitted Gaussians are simply to aid visualization of the distributions' separability.

It was also observed that the accuracy of the predictions increased when the classifier was trained for a particular person. We believe this is important because people have different speaking styles, which potentially could be captured by our model. We had 8 conversations involving subject J and trained our model only on her data set. For the romantic interest question for subject J, for example, the cross-validation accuracy of our predictions increased to 87.5%.

We also used SVM classifiers to test the separability of the 'yes' and 'no' responses, with linear, polynomial and radial kernels. As seen, the performance of the SVM classifier was only slightly better than a simple linear classifier (with appropriate feature selection). The RBF kernel is more prone to overfitting.

Question	SVM accuracy (Linear kernel)	SVM accuracy (RBF kernel)
Are you interested in this person? (asked of females)	0.71	0.70
Would you like to be friends with this person? (asked of females)	0.76	0.79
Are you interested in maintaining a professional relationship with this person? (asked of females)	0.73	0.71
Are you interested in this person? (asked of males)	0.64	0.62
Would you like to be friends with this person? (asked of males)	0.79	0.82
Are you interested in maintaining a professional relationship with this person? (asked of males)	0.71	0.69

2.4 Real-time Feedback

If our description of social signaling is precise, and the effects are measurable, then a wearable 'social signaling meter' is the next step. We wanted to use such a device to understand how speed daters would react if made more aware of their social signals.

We have built a real-time speech feature calculation engine (SpeedDating Pro) that runs on a Sharp Zaurus Linux PDA. The Zaurus is a wireless, ARM-based handheld with a microphone input and a touch screen interface. The current version of the speech feature code is written in C++, and incorporates all features except the engagement features (or the influence parameters). It also calculates the probability of the other person saying 'yes' based on the coefficients from the model training. Zaurii are time synced using ntpdate (a linux utility) and use UDP broadcasts to communicate with each other. The display can show chances of a second date or the measured social signals, and is updated after every 90 seconds of conversation.



Figure 4 (a) (left): Display of probability of the other person saying YES



Figure 4 (b) (right): Display of measured social signals

We are deploying 20 such Zaurus PDAs running SpeedDating Pro software soon, and look forward to the results. SpeedDating Pro is in some ways similar to the 'Love Detector' software sold by Nemesysco, although we were unable to find any publications documenting its accuracy. We believe that the underlying software for Nemesysco is based on vocal stress features, and in our preliminary trials with it, it had ambiguous results.

3 Future Work

We believe that these same social signals may also help us to predict whether people are interested or engaged in a much broader range of situations. In one study that is just beginning to be analyzed, we have people talk to each other on various topics, and then compare the measured signals to their interest rating on a feedback form. Based on the results of the first 8 subjects, which were engaged in 80 conversations of about 3 minutes each, there is a correlation of $r=0.64$, $p=0.1$ between the social signalling and the rated interest level.

Annual US spending on focus groups for qualitative market research is about US\$ 1.1 Billion. However, industry experts acknowledge that focus groups are largely lacking in accuracy, since participants often misreport experiences and say what is expected of them, without sharing their insight as a consumer. By measuring these non-verbal signals, the efficacy of focus groups could potentially be improved.

4 Discussion

The social signalling measured in this paper seems to communicate and be involved in mediating social variables such as status, interest, determination, or cooperation, and arise from the interaction of two or more people rather than being a property of a single speaker. Semantics and affect are important in determining what signalling an individual will engage in, but they seem to be fundamentally different types of phenomena. The social signalling measured here seems to be a sort of 'vocal body language' that operates relatively independently of linguistic or affective communication channels, and is strongly predictive of the behavioural outcome of dyadic interactions.

5 Acknowledgements

The authors would like to express their gratitude to Ashish Kapoor and Rakhi Bhavnani. We would also like to thank our UROP engineering team of Shaun Foley, Harvey Jones, Mike Tully Klein Jr. and Kevin Kim.

6 References

- [1] Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behaviour as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2), 256-274.
- [2] Basu, B., (2002) Conversational Scene Analysis, doctoral thesis, Dept. of Electrical Engineering and Computer Science, MIT. 2002. Advisor: A. Pentland
- [3] Breazal C. (2001), *Designing Social Robots*, MIT Cambridge Press
- [4] Chartrand, T., and Bargh, J., (1999) The Chameleon Effect: The Perception-Behavior Link and Social Interaction, *J. Personality and Social Psychology*, Vo. 76, No. 6, 893-910
- [5] Choudhury, T., and Pentland, A., (2004), NAASCOS, June 27-29, Pittsburgh, PA. PDF available at <http://hd.media.mit.edu>
- [6] Fernandez, R. (2004), A Computational Model for the Automatic Recognition of Affect in Speech, Ph.D. Thesis, MIT Media Arts and Science, February 2004. Advisor: R. Picard
- [7] Gladwell Malcolm (2004) , *Blink*, Little Brown and Co. NY 2004
- [8] Handel, Stephen, (1989) *Listening: an introduction to the perception of auditory events*, Stephen Handel , Cambridge: MIT Press
- [9] Jaffe, 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*, 66(2), No. 264.
- [10] Khilnani, R. (2004) Temporal Analysis of Stages in Negotiation, MEng Project, Advisor: A. Pentland.
- [11] Kurzban, Weeder, to be published in journal of *Evolution and Human Behaviour*
- [12] Nass, C., and Brave, S. (2004) *Voice Activated: How People Are Wired for Speech and How Computers Will Speak with Us*, MIT Press
- [13] Oudeyer, P-Y. (2003) The production and recognition of emotions in speech: features and algorithms. *International Journal of Human Computer Interaction*, 59(1-2):157--183.
- [14] Picard R. (1997) , *Affective Computing*, MIT Press
- [15] Pentland, A. (2004) *Social Dynamics: Signals and Behavior*, ICDL, San Diego, CA Oct 20-23, IEEE Press
- [16] Pentland, A., Curhan, J., Khilnani, R., Martin, M., Eagle, N., Caneel, R., Madan A (2004) ``Toward a Negotiation Advisor," UIST 04, Oct 24-27, ACM. PDF available at <http://hd.media.mit.edu>