Visualisation and Prediction of Conversation Interest through Mined Social Signals

Dumebi Okwechime, Eng-Jon Ong, Andrew Gilbert, and Richard Bowden
CVSSP, University of Surrey, Guildford, Surrey, GU17XH, UK
{d.okwechime, e.ong, a.gilbert, r.bowden}@surrey.ac.uk

***Abstract*— This paper introduces a novel approach to social behaviour recognition governed by the exchange of non-verbal cues between people. We conduct experiments to try and deduce distinct rules that dictate the social dynamics of people in a conversation, and utilise semi-supervised computer vision techniques to extract their social signals such as *laughing* and *nodding*. Data mining is used to deduce frequently occurring patterns of social trends between a speaker and listener in both *interested* and *not interested* social scenarios. The *confidence* values from rules are utilised to build a Social Dynamic Model (SDM), that can then be used for classification and visualisation. By visualising the rules generated in the SDM, we can analyse distinct social trends between an *interested* and *not interested* listener in a conversation. Results show that these distinctions can be applied generally and used to accurately predict conversational interest.**

I. INTRODUCTION

As naturally social entities, humans can easily extract social information from non-verbal communication without the need of understanding what is being said. Psychologists believe this skill is hard-wired in the human brain [1]. Gesture, vocal signal, and body language triggers unconscious analysis of socially relevant information [2]. Since non-verbal communication plays such an important role in our social interaction, a method of modelling it would prove valuable in understanding relationships, identifying context/intent, or generating synthetic responses in an Artificial Intelligent (AI) context.

Our aim is to devise a model for non-verbal communication. It will allow classification and visualisation of multimodal exchanges in social signals between a speaker and listener in a conversation. Unlike other social models that rely on intangible psychological observations, we propose the use of tangible rules governed by the data to discern distinct trends and characteristics. We achieve this by using data mining [3] to efficiently identify social trends between a group of three people in a conversation. The *confidence* values are used to build a *Social Dynamics Model (SDM)*. SDM allows for efficient visualisation of multimodal social trends, to enable visual distinctions between the two scenarios. Using these distinctions, the modal accurately predicts conversational interest within a window of less than five minutes.

This work was supported by the EPSRC project LILiR (EP/E027946/1) and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 231135 - DictaSign.

This paper is divided into the following sections. Section II briefly details a background in different approaches for understanding social interaction. Section III presents our dataset and methods for conversational analysis. Section IV describes the visualisation of the SDM, and the remainder of the paper presents an evaluation and conclusions.

II. BACKGROUND

Traditional social interaction research can be grouped into two main categories: emotion, based on cognitive psychology [4], and linguistics, based on dialogue understanding [5][6]. Although emotion understanding is of vital importance in how people socially interact, emotion recognition in a natural conversation is a very complex problem and would require extensive data and research in deducing social trends. Also, structured dialogue can not be easily interpreted to observe generalised behaviour. Methods that utilise machine learning models such as HMMs [7] or Dynamic Bayesian Networks [8], are applied to generic features in audio signals and pixel intensities to discern social behaviour. However, in this paper we derive rules of social behaviour by focusing mainly on non-verbal social cues.

Psychological studies have proven that observing non-linguistic/non-verbal, unconscious social signals [9], can provide effective information in social interaction understanding [1]. A number of researchers have used machine analysis of non-verbal social signals to interpret social behaviour. The idea of Social Signal Processing [10][11], originally introduced by Pentland [12], is to use visual and vocal analysis to understand social behaviour and predict outcomes of dyadic interactions to enable a *Human-Centred* computing paradigm. This is achieved using *textures* (i.e. speaker energy and amount of movement) [13] from multimodal social signals. Similarly, Curhan et al [14] use these *texture* features to predict outcomes of negotiations based on 'thin slices' [15] of employment negotiation data. Although these methods perform well in predictions, they rely on psychological observations to derive prior assumptions of what is positive or negative social behaviour. This may not be accurate in all social contexts. Also, their approach is unable to discern co-occurence of social signals of multiple modes, as these more complex dependencies are difficult to identify. However, in this work, we introduce the SDM, which utilises data mining to derive tangible rules for visualising multimodal social interaction and for accurately predicting social context. This

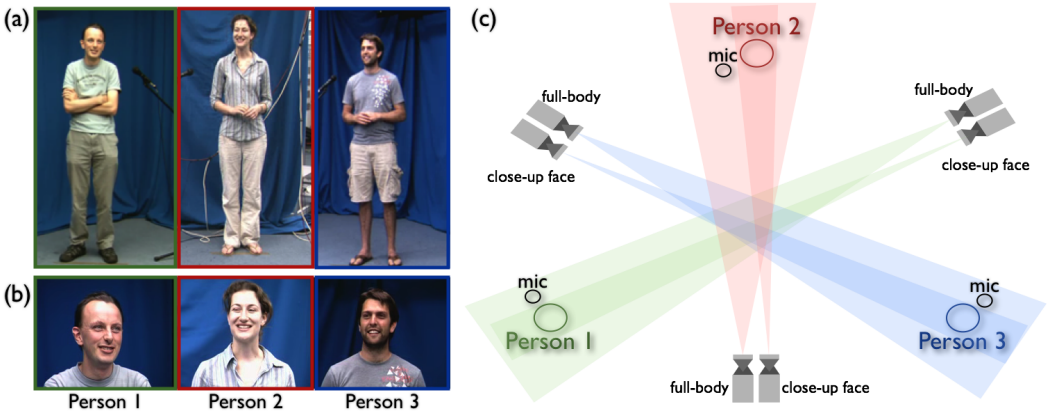

Fig. 1. (a) Image showing full-body view of recorded video data of three people having a conversation. (b) Image showing close-up face view. (c) Diagram showing the configuration of cameras, microphones (mic) and conversers. We refer to the three people in the conversation as person 1, 2, and 3.

is achievable independent of prior psychological evaluations, relying solely on the trends in the data.

Eagle et al [16] introduce *reality mining* which uses mobile devices, like smart badges and cellular phones, to extract proximity and vocal information to derive social networks. The main difference to our method is we use multimodal social signals as features for apriori association rule mining, with the aim of deriving specific association rules that govern conversational interest.

III. CONVERSATION ANALYSIS

A. Dataset

The dataset [1] is composed of approximately 30 minutes of audio and video recordings (43000 frames) of the full-body frontal view (516×340, 25 frames per second, 48kHz) and the close-up face view (720×576, 25 frames per second, 48kHz) of 3 individuals having a conversation with each other. An image of the full-body recording for each person is shown in Figure 1(a), and the face recording in Figure 1(b). We refer to the 3 individuals as person 1, 2, and 3. Each person remained in a stationary position relative to the cameras as shown in Figure 1(c).

Prior to capture, each person was given a questionnaire and asked to score from 1-3 their interest (where 1 is low interest and 3 is high interest) on a given set of *book genres*, *film genres* and *music genres*. They were also given specific questions like: *favourite sports*, *language(s) spoken fluently*, *favourite music concerts*, *favourite theatre shows* etc. Their questionnaires were analysed to choose topics for conversation that would lead to the following 4 generic scenarios:

- All interested in topic
- Two people interested in topic, one person is not
- One person interested in topic, two people are not
- None are interested in topic

These 4 generic scenarios where derived from 8 topics of conversation as detailed in Table I. The sixth column of Table I shows the limited duration of each topic, chosen to suite the scenario. A projector displayed the topic of conversation for discussion, and a quiet bell would ring to make the subjects aware of the change in topic. The subjects were unaware of the nature of the experiment, and were simply asked to discuss the topic displayed on the screen.

The aim of this experiment was to observe the social dynamics between the three people in scenarios when *interested* or *not interested* in the topics. The next step is to quantise their social signals (from the data) in a form that is suitable for data mining in order to obtain rules governing social behaviour.

B. Social Signals

Pentland [13] proposed measuring non-linguistic social signals using four main observations: *activity level*, *engagement*, *emphasis* and *mirroring*. Using this as our base, we chose to observe 7 social signals in the conversation: *Voiced*, *Talking*, *Laughing*, *Head Shake*, *Head Nod*, *Activity Measure*, and *Gaze Direction*.

We use a variety of techniques to derive each of these labels.

1) **Voiced[V]:** The audio stream is represented using 12 MFCCs (Mel-Frequency Cepstral Coefficients) and a single energy feature of the standard HTK setup [17]. For each person, a few voiced segments were labelled and a Mahalanobis distance measure was used to derive a correlation between the voiced and non-voiced regions.
2) **Talking[T]:** With the voiced segments labelled, it was a simple process of labelling the voiced segments which were talking. This was done by hand.
3) **Laughing[L]:** The Viola-Jones face detector [18] was used to segment the face region in each frame. The lip region was localised by cropping the lower-centre region of the face. An AdaBoost classifier was then trained for laughing and used to label the remaining data.
4) **Head Shake[S]:** The Viola-Jones face detector was used to determine the movement of the face. A Fast

[1]The dataset along with annotation can be made available upon request. Please email {d.okwechime@surrey.ac.uk}.

Scenario	P1	P2	P3	Topic	Period
1	3	3	3	Classical Music	5 min
2	3	3	1	Adventure Novels	5 min
3	3	1	3	Philosophy Novels	5 min
4	1	3	3	Rock Music	5 min
5	3	1	1	Sailing (Spoken in French)	2.5 min
6	1	3	1	Triathlon/Les Miserables (Spoken in Afrikaans)	2.5 min
7	1	1	3	Radio Head Concert	2.5 min
8	1	1	1	Horror Novels	1.5 min

TABLE I

TABLE SHOWING 8 DIFFERENT SOCIAL SCENARIOS DICTATED BY THE TOPIC OF CONVERSATION. THE THREE PEOPLE ARE REFERRED TO AS **P1** FOR PERSON 1, **P2** FOR PERSON 2, AND **P3** FOR PERSON 3. THE NUMBERS INDICATE THEIR INTEREST IN THE TOPICS WHERE 3 IS A HIGH INTEREST AND 1 IS A LOW INTEREST

Fourier Transform (FFT) was used to identify high frequency movement along the x-axis

5) **Nod[N]:** Similar to head shakes, an FFT was used to identify high frequency movement along the y-axis.
6) **Activity Measure[A]:** The torso region of the full body video was segmented using colour and the mean-scaled standard deviation of velocity was measured. The leg and head regions are ignored because, there was minimal leg movement (subjects remained stationary), and since we are more interested in gesture activity, changes in head posture/gaze would bias the activity measure.
7) **Gaze Direction[G]:** The eye pupils and the corners of the eyes were tracked using a Linear Predictor tracker [19]. The corners of the eyes were normalised to 0 and 1, and the position of the eye pupil within this region was used to determine if the person was gazing left [GL], right [GR] or centre [GC].

This produces N_T sets of social signal labels (where N_T is the total number of frames) of 27 dimensions, where $1-9$ is for person 1, $10-18$ for person 2 and $19-27$ for person 3. We define 2 complete sets of social signal vectors for *interested* and *not interested* scenarios as $F_{(Int)}$, and $F_{(NotInt)}$ such that $F = \{\mathbf{f}_i\}_{i=1}^{N_T}$ where $\mathbf{f}_i$ is a 27 dimensional binary vector.

C. Mining for Frequent and Distinctive Social Trends

This experiment is driven by the speaker. At any given time, there is only one speaker and one listener. We are interested in the combination of social signals a listener performs when *interested* and *not interested* in the conversation. Manually observing all combinations of listener and speaker behaviours in such a large data set would be virtually impossible. A solution would be to make some common sense prior assumptions of expected trends (i.e. an interested listener would gaze more at the speaker than when they are not interested) and focus primarily on these assumptions. However, there is no way of proving or disproving such assumptions, and, there is a large list to chose from.

We wish to employ a data driven approach to learn such rules. We propose a novel approach to deriving social dynamics and trends between the subjects based on data mining [3]. Data mining allows for large data sets to be processed to identify the reoccurring patterns within the data in an efficient manner. In this work, Apriori Association rule [3][20] mining is used. Formally developed for supermarkets to analyse millions of customer's shopping trends, we aim to find *association rules* between a speaker and listener that indicate *interested* and *not interested* from the multitude of possible rules that could exist.

An association rule is a relationship of the form $\{R_i^A\} \Rightarrow R_i^C$ where R_i^A is a set of social signals of the speaker, and R_i^C a social signal of the listener. $R_i^A = \{r_{i,1}^A, ..., r_{i,|R_i^A|}^A\}$ is the antecedent where r_i^A denotes a speaker's social signal, and $R_i^C = \{r_{i,1}^C, ..., r_{i,|R_i^C|}^C\}$ the consequence where r_i^C is a listener's social signal. An example would be, if $R_1^A = \{[N],[L]\}$, and $R_1^C = \{[L]\}$ as defined in Section III-B, then, $\{R_1^A\} \Rightarrow R_1^C$ would imply "when the speaker nods and laughs, the listener is very likely to also laugh". The belief of each rule is measured by a *support* and *confidence* value. The *support* measures the statistical significance of a rule, it is the probability that a transaction contains itemset R_i^A.

$$sup(\{R_i^A\} \Rightarrow R_i^C) = sup(\{R_i^A\} \cup R_i^C) \tag{1}$$

The *confidence* is the number of occurrences in which the rule is correct, relative to the number of cases in which it is applicable.

$$conf = \frac{sup(\{R_i^A\} \cup R_i^C)}{sup(R_i^A)} * 100 \tag{2}$$

Apriori Association mining is applied to the social signal labels for both *interested* listener and *not interested* listener scenarios, to derive frequently occurring association rules.

Traditionally, data mining looks for a combination of symbols that occur simultaneously. However, a listener's social behaviour is always a response to the speaker's social signals, hence, co-articuation is not possible. To account for this, *temporal bagging* within a set temporal window is used to enforce a temporal coherence between features. Given a speaker's social signal, we observe the listener's social behaviour $s = 10$ frames in the future (approx $\frac{1}{2}$ a second).

IV. VISUALISING AND INTERPRETING SDM

The SDM allows visualisation of multimodal trends in social interaction between a speaker and listener in a conversation. Using the mined confidence values, the conditional probability of the listener's social responses given the

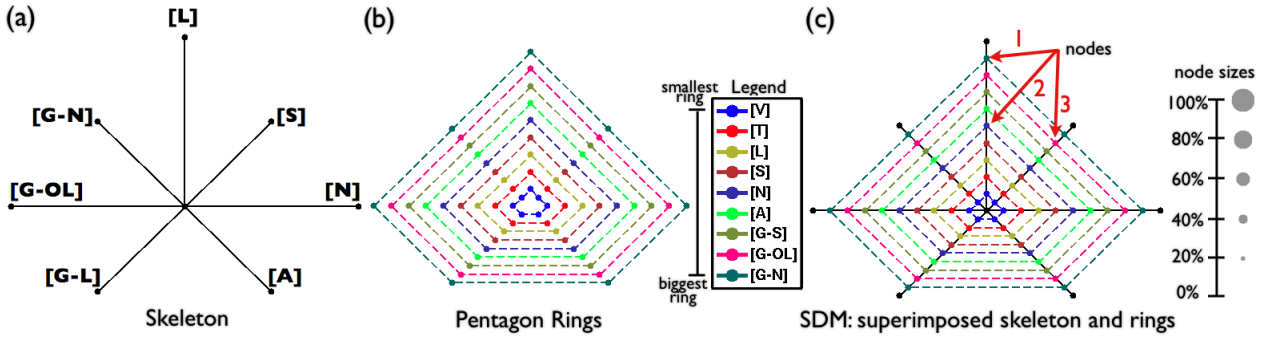

Fig. 2. (a) The skeleton of SDM. Consists of 7 black lines that are attached to a central intersection. Each line represents a different speaker's social signal (b) 9 pentagon rings where each ring represents a listener's social response. The individual rings are coded by colour and size. (c) SDM is made up of the rings superimposed on the skeleton. The points where the rings intersect the skeleton are known as nodes and infer a listener's social response given a speaker's social signal. A few are indicated by the three red arrows (arrows 1, 2, 3). Arrow 1 is pointing at node [L] ⇒ [G-N], arrow 2 at node [L] ⇒ [N], and arrow at node [S] ⇒ [G-OL]. The idea is that the nodes will vary in size reflecting the respective mined confidence value.

speaker's social signals, can be observed efficiently to distinguish social trends without needing to rely on psychological observation.

To avoid over complicating the diagram with the numerous combinations of association rules, to visualise the SDM, only association rules with single antecedents (i.e. $|R_i^A| = 1$) are used, whereby the likelihood of a listener's social response is derived by a single speaker's social signal. The more complex rules are still kept in the model, however, we use the simpler rule for visualisation to discern prominent trends.

A. Interpretation of an SDM

The SDM is made up of two components: a **skeleton** and a set of **pentagon rings**.

The skeleton consists of 7 black lines that collectively meet at a central intersect. Each line represents a different speaker's social signal and are configured as shown in Figure 2(a) (the labels are as detailed in Section III-B). With regards to the speaker, *voiced* ([V]) and *talking* ([T]) social signals are ignored since the speaker is guaranteed to be voiced and talking. To add clarity to the gaze labels, instead of [GL], [GR] and [GC], we use [G-S], [G-OL], [G-N] representing *gazing at speaker*, *gazing at another listener* and *gazing at no one*, respectively. This allows us to know who they are gazing at.

The second component is a set of 9 pentagon rings. Each ring represents a different listener's social signal. As presented in Figure 2(b), the individual rings are coded by colour and size. Shown in Figure 2(c), the superimposed skeleton and pentagon rings make up the SDM. The points where the rings intersect the skeleton infer the occurrence of a listener's social response given a speaker's social signal. We refer to these points as nodes, three of which are indicated by the red arrows (arrows 1, 2, and 3) in Figure 2(c). Arrow 1 is pointing at node [L] ⇒ [G-N], denoting that when the speaker laughs, the listener gazes at no one. Arrow 2 is pointing at node [L] ⇒ [N] (when the speaker laughs, the listener nods), and arrow 3 at node [S] ⇒ [G-OL] (when the speaker shakes their head, the listener gazes at another listener). These nodes can vary in diameter, reflecting the size of the mined *confidence* value given the rule. A set of example node sizes are presented in the right corner of Figure 2(c). Using this structure, we can efficiently visualise prominent rules when comparing social scenarios, simplifying a potentially complex set of social behavioural information.

V. EVALUATION

A. IDENTIFY DISTINCT TRENDS

To identify trends in social behaviour between a listener and speaker in an *interested* and *not interested* scenario, we performed data mining separately on our *interested* and *not interested* datasets of social signal labels (as detailed in Section III). 357 rules in total were extracted from the mining in the *interested* scenario, 63 of which had single antecedents (i.e. $|R_i^A| = 1$), 153 with two antecedent, 133 with three antecedents, and 8 with four antecedents. Mining in the *not interested* scenario extracted 396 rules, consisting of 63 rules with single antecedents, 162 with two antecedents, 162 with three antecedents, and 9 with four antecedents. Such complex rules (up to 4 dimensions/antecedents) would be impossible to derive any other way than analytically. Using the *confidence* values derived from these rules, two SDMs were built as shown in Figure 3(a) and 3(b). Figure 3(a) is the SDM of a speaker given an *interested* listener and Figure 3(b) is the SDM of a speaker given a *not interested* listener.

By observing both diagrams, the similarities they share are instantly noticeable. All nodes on the third pentagon from the top (third biggest ring), representing the listener's social response [G-S] (gazing at speaker). These are prominent in all instances of the speaker's social signals in both diagrams. A similar trend exists (with minor variations) in nodes on the third pentagon from the bottom (third smallest ring), representing the listener's social response [L] (laughing). From this observation, we can see that contrary to some social interaction studies, neither a constant gaze at speaker nor long periods of laugher, can distinguish between an interested or not interested listener in a conversation.

The clearest distinction between the two diagrams are the nodes on the smallest pentagon (colour coded light blue) representing the listener's social response [V] (voiced). Voiced regions imply an exchange of short single words like

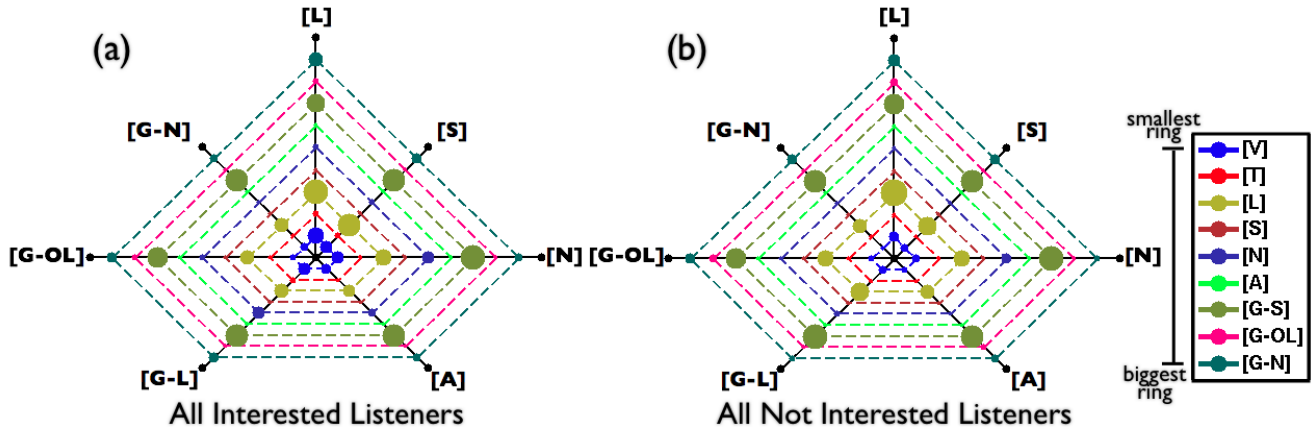

Fig. 3. (a) SDM generated from the mined *interested* listener's confidence values. (b) SDM generated from the mined *not interested* listener's confidence values.

'uh-huh' or 'yea'. Voiced [V] is a vocal form of *backchannel response* [21][22]. *Backchannel responses* are used by the listener to give feedback to the speaker, expressing acknowledgement, understanding, and presence in the conversation [23]. Here we see that a majority of these [V] nodes are bigger in the interested scenario when compared to the not interested scenario, especially in the rule [G-L] $\Rightarrow$ [V], when the speaker is gazing directly at the listener.

The next discernible trend belongs to the nodding nodes [N] on the fifth pentagon from the centre (colour coded dark blue). Similar to voiced [V], nodding [N] is a visual form of *backchannel response*, used to confirm engagement in the conversation. Although the listener in the *not interested* scenario mirrors the speaker well in comparison to the *interested* scenario (i.e. [N] $\Rightarrow$ [N]), the listener in the *not interested* scenario barely nods in response to any other social signals. In this case, mirroring is not a discerning social behaviour between an *interested* and *not interested* listener. However, from the diagram we can see that the *interested* listener nods more consistently in response to the other social signals, especially when the speaker gazes directly at the listener (i.e. [G-L] $\Rightarrow$ [N]).

The final discernible trend is the talking social response [T] (second pentagon from the centre, colour coded red). While only a mild occurrence in the *interested* scenario, it rarely occurs at all in the *not interested* scenario. Talking [T] suggests *turn-taking* [24], whereby the listener attempts to participate in the conversation whilst the speaker is speaking.

To validate these findings, using these confidence values, we take the ratio of *interested* and *not interested* results for matching rules. We obtain results of greater than 1 when the rule occurs more frequently in the *interested* scenario, and less than 1 when they occur more in the *not interested* scenario. The results are shown in Table II. Rows 1, 2, and 5 relating to the listener's social responses [V], [T], and [N] respectively, are the most prominent distinct trends between the two scenarios, with an average of greater than 1.5, as shown on the last column of Table II. Also, rows 3 and 4 relating to the listener's social responses [L] and [G-S], produce an average of 1, varying equally in both the interested and not interested social scenarios. This is analogous to our earlier observations.

This proves that a clear distinction between an *interested* and *not interested* listener can be determined using the SDM. The next step is to use these distinct social signals for conversation interest prediction.

B. CONVERSATION INTEREST PREDICTION

Using the discerning conversation social signals, we attempt to use the SDM to accurately predict conversation interest. To perform this test, we eliminate one person's social activity from our dataset, then perform mining using only the other two people's social interaction (with each other). This is done for the combined dataset of the *interested* and *not interested* scenarios, resulting in an SDM. This SDM becomes the trained classifier. We then observe all the eliminated person's social responses when in the role of a listener in both social scenarios. Using the SDM classifier, we attempt to predict conversation interest based on the generalisation of rules across the subjects the model was trained on. The predictions are done on the entire dataset using different time frame windows ranging from 100 frames (4 seconds) to 7000 frames (approx $4\frac{1}{2}$ minutes) with 1 frame increments. There are three people in our dataset, so we are able to perform this test three times (once for each person), alternating the eliminated listener. The results are shown in Figure 4.

As shown in Figure 4, with only 4 seconds of observation, we obtain predictions better than random. However, as more evidence accumulates, the performance increases to 90% for a time window of approximately $4\frac{1}{2}$ minutes. Theses results prove the SDM can derive distinct social trends between the two scenarios, which can generalise well for accurate predictions.

VI. CONCLUSION

SDMs can accurately predict conversational interest. Unlike other methods, we show that by using data mined *confidence* values, discerning trends in exchanges in social signals is straight forward, without the need for psychological observations. This approach is not context dependent

		Speaker							
		[L]	[S]	[N]	[A]	[G-L]	[G-OL]	[G-N]	Aver
Listener	[V]	**1.4**	**1.7**	**1.5**	**1.4**	**1.5**	**1.1**	**1.7**	**1.5**
	[T]	**2.2**	**3.9**	**14**	**2.2**	**2.7**	**2.5**	**4.4**	**4.6**
	[L]	0.9	1.2	1	1	0.8	0.9	1	1
	[S]	0.5	0.7	5	0.7	0.7	0.3	0.6	1.2
	[N]	**1.8**	**2**	**1.2**	**1.6**	**1.9**	**1.6**	**1.9**	**1.7**
	[A]	0.8	1.2	1.3	1.6	1.1	0.9	1.1	1.1
	[G-S]	0.9	1	1	1	0.9	1	1	1
	[G-OL]	0.8	0.9	0.7	0.7	1	0.7	1	0.9
	[G-N]	1.2	1.1	1.3	1.1	1.5	1	0.9	1.1

TABLE II

RATIO OF *interested* TO *not interested* CONFIDENCE VALUES FOR MATCHING RULES. AN AVERAGE IS CALCULATED IN THE LAST COLUMN.

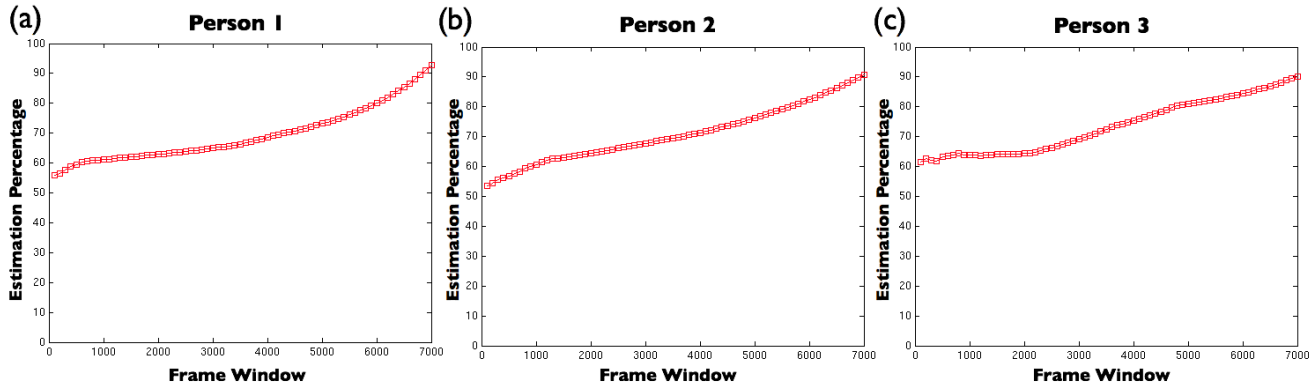

Fig. 4. Prediction percentage scores using varying frame windows for each person.

and future work will explore larger datasets of other social scenarios such as debates, arguments and social signal exchanges between partners. A means of real-time predictions would be a valuable addition, which only requires a tool for extracting social signals of multiple people from a live video stream.

One of the main benefits of this model is it has vast applications. It can be used for generating social behaviour in computer generated avatars used in the movie and gaming industries and in Human Computer Interaction (HCI). It can also be used to assist social scientists and medical psychologists in diagnosing certain social related conditions with just a short period of observation.

REFERENCES

[1] M. Knapp, "Nonverbal communication in human interaction," in *Harcourt Brace College Publishers*, 1972.

[2] M. Argyle, "The psychology of interpersonal behaviour," in *Penguin*, 1967.

[3] A. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *In Proc. of the ACM SIGMOD Int. Conf. on Management of Data SIGMOD*, 1993.

[4] P. Ekman and W. Friesen, "Facial action coding system," in *Consulting Psychologists Press, Palo Alto, CA*, 1977.

[5] M. Argyle, "Bodily communication," in *Methuen*, 1987.

[6] A. Kendon, R. Harris, and M. Key, "Organisation of behavior in face to face interaction," in *The Hogue, Netherlands: Mouton*, 1975.

[7] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio, "Detecting group interest-level in meetings," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2005.

[8] A. Dielmann and S. Renals, "Automatic meeting segmentation using dynamic bayesian networks," *IEEE Trans. on Multimedia*, 2007.

[9] N. Ambady, F. Bernieri, and J. Richeson, "Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream," *Advances in experimental social psychology*, 2000.

[10] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[11] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social signal processing: state-of-the-art and future perspectives of an emerging domain," in *Proceeding of the 16th ACM Int. Conf. on Multimedia*. ACM, 2008, pp. 1061–1070.

[12] A. Pentland, "Social signal processing." in *IEEE Signal Processing Magazine*, 2007, pp. 108–111.

[13] ——, "A computational model of social signaling," in *18th Int. Conf. on Pattern Recognition. ICPR*, 2006.

[14] J. Curhan and A. Pentland, "Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes." *Journal of Applied Psychology*, vol. 92, no. 3, pp. 802–811, 2007.

[15] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis." *Psychological Bulletin*, vol. 111, no. 2, pp. 256–274, 1992.

[16] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal and Ubiquitous Computing*, 2006.

[17] A. Mertins and J. Rademacher, "Frequency-warping invariant features for automatic speech recognition," in *2006 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2006.*, vol. 5, 2006.

[18] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," in *Proc. IEEE CVPR*, 2001.

[19] E. J. Ong, Y. Lan, B. J. Theobald, R. Harvey, and R. Bowden, "Robust facial feature tracking using selected multi-resolution linear predictors," in *Int. Conf. Computer Vision*, 2009.

[20] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *In VLDB'94 Proc. of 20th Int. Conf. on Very Large Data Bases*, 1994, pp. 487–499.

[21] V. Yngve, "On getting a word in edgewise," in *Papers from the sixth regional meeting of the chicago linguistic society*, 1970, pp. 567–577.

[22] A. Mulac, K. Erlandson, W. Farrar, J. Hallett, J. Molloy, and M. Prescott, "'uh-huh. what's that all about?' differing interpretations of conversational backchannels and questions as sources of miscommunication across gender boundaries," in *Communication Research*, 1998.

[23] M. Schröder, D. Heylen, and I. Poggi, "Preception of non-verbal emotional listener feedback," in *Proc. of Speech Prosody*, 2006.

[24] E. Goffman, "Replies and responses," in *Language in Society*, 1976.