ARE YOU A WEREWOLF? DETECTING DECEPTIVE ROLES AND OUTCOMES IN A CONVERSATIONAL ROLE-PLAYING GAME

Gokul Chittaranjan[1], *Hayley Hung*[2]

[1]Department of Electrical and Computer Engineering, University of Florida, Gainesville 32611, USA
[2]Idiap Research Institute, P.O. Box 592, CH-1920 Martigny, Switzerland

ABSTRACT

This paper addresses the task of automatically detecting outcomes of social interaction patterns, using non-verbal audio cues in competitive role-playing games (RPGs). For our experiments, we introduce a new data set which features 3 hours of audio-visual recordings of the popular "Are you a Werewolf?" RPG. Two problems are approached in this paper: Detecting lying or suspicious behavior using non-verbal audio cues in a social context and predicting participants' decisions in a game-day by analyzing speaker turns. Our best classifier exhibits a performance improvement of 87% over the baseline for detecting deceptive roles. Also, we show that speaker turn based features can be used to determine the outcomes in the initial stages of the game, when the group is large.

Index Terms— Deception, Role Analysis, Nonverbal Behavior

1. INTRODUCTION

Recently, there has been a growing interest in the analysis of non-verbal social signals using machine learning and signal processing by the computer science community [1, 2, 3]. The understanding of behavior in teams, that can be either cooperative or competitive, is of vital importance for organizational management [4]. Understanding individual behavior and estimating when people may be suspicious of others is important for preventing the breakdown of trust within teams or other social relationships.

Thus the objective of this work is to discover through a social setting, how deceptive roles affect the way people behave both as actors and perceivers of a particular behavioral type. In addition, since we are considering deception in the context of role playing games, questions are raised about why people are willing to become engaged even when rewards are small. One could conclude that the thrill of living a different role that contrasts with our daily lives can help us to de-stress. Understanding the mechanisms in games that encourages enjoyment and escapism can also help in the design of better games. Finally, automatic estimates of roles and deceptive behavior in the context of game playing can be useful for mining and browsing personal recordings of families and friends as storage and recording devices become a standard commodity but organizing and finding past events becomes more difficult.

In social psychology, the analysis and detection of deception is an active research area [5] for both crime prevention and psychiatric purposes. Deception is defined as "a successful or unsuccessful deliberate attempt, without forewarning, to create in another a belief which the communicator considers to be untrue" [5] (p.15). Detecting deception can be approached from two perspectives; we can analyze verbal responses, or, alternatively, non-verbal behavior [5]. This can be inferred from vocal and/or visual cues [5]. Vocal cues refer to pitch, pauses, laughter etc. and visual cues refer to illustrators[1], body movements, eye blinks etc. Of these, increase in pitch and decrease in illustrators, have been found to be the strongest indicators of deception with d=0.2 and -0.36 respectively (see Section 6) [5]. However, these values depend on several factors such as the motivation, the stakes involved and the personality of the liar [5].

Recently, automatic detection of lies using machine learning has been investigated. Hirschberg et al. [6] introduced the Columbia-SRI-Colorado (CSC) Corpus that contains audio with transcription of interviews of participants, who are motivated to lie for financial and self-presentational gains. It shows that acoustic/prosodic cues give a slightly improved performance, with an error of 38.5% over the baseline of 39.8%, in which the most probable class (truth) is chosen all the time. Another study on the CSC corpus [7], shows that a combination of prosodic, acoustic and lexical features drops the error to 36%. Also, T.O. Meservy et. al [8] investigated supervised learning of visual cues for detecting lies. All the work described above had liars face interviews, in a dyadic scenario.

In another study, M. Sung and A. Pentland describe the use of non-invasive sensors to gather physiological and acoustic data in a poker playing scenario [9]. The cues used in this work, resemble those used in polygraph examinations [5]. However, the this method is obtrusive. To our knowledge, no work has been done on the more challenging problem of automatically detecting deception in large social groups using non-invasive methods.

Automatically estimating roles has been tested on televised competitions/debates and cooperative meetings [2, 3]. S. Favre et. al. [3] used televised political debates where two teams of two choose a position to debate on, so those on opposing teams would tend to talk after each other. It was shown that affiliation networks built from speaker turns could be used for detecting roles. In some cases [2], an open and competitive environment is present, in which participants freely interact with each other. In [2], features derived from speaker turns were used with a ranking-based scheme, to predict outcomes.

Here, we present work on two different tasks: (i) the use of a relevance vector machine that uses non-verbal audio cues to automatically detect liars in a role playing game; (ii) automatically predicting decisions made in the game by analyzing speaker turns.

The paper is organized as follows: Section 2 describes the data set. Section 3 defines the two tasks of deception/suspicious behavior detection and predicting outcomes in the game. Section 4 deals with the feature extraction. Section 5 describes our approach in addressing the two tasks. Section 6 presents the results. Finally, Section 7 enlists conclusions and guidance for future work.

The work of G. Chittaranjan and H. Hung has been supported by European IST Program Project FP6-033812 (AMIDA)

[1]hand/arm movements, that modify and/or supplement what is being said verbally

2. THE "WEREWOLF" DATA SET

"Are you a Werewolf?" is an RPG suited for large groups and is a game of accusations, lying, second-guessing, assassination and mob hysteria [10]. This game was particularly chosen for three reasons: (i) deception occurs in a large social group where the liars are an informed minority and must be found through discussions; (ii) collective decisions are made through discussions about who the liars are, which enables us to easily record decisions made by the players; (iii) deceptive and suspicious behavior can be differentiated easily. A detailed mathematical analysis of the outcomes of this game, for different group sizes has been done in the past [11].

The game proceeds as follows: The players are divided into villagers and werewolves. Some villagers also have special roles (explained later). The game proceeds in two alternating phases. The game-night, in which the villagers are asleep (have their eyes closed) and the werewolves kill a villager of their choice and the game-day in which everybody is awake. In this phase, all players can talk and decide to lynch a person whom they believe to be a werewolf. Once someone is lynched, they are out of the game and may not speak. The game ends when all werewolves are dead (villagers win) or the number of werewolves and villagers become equal (werewolves win).

Many variants of the game exist [10]. Our recordings contain 2 werewolves per game, with a seer and/or a little girl. The choice of whether these roles existed in a game as well as the assignments were made at random. The seer can obtain the real identity of any person during each night phase. The little girl is active all the time and can sneak-peek at what happens in the night. The seer and little girl are "special" roles assigned to certain villagers who must be careful not to reveal their identity too early on in the game, for fear of being killed by the werewolves.

The "Werewolf" data set consists of audio-visual recordings of 8 games played by 2 groups of people. In total 18 different people were recorded for the data set. The groups had a size of 10 and 8 participants respectively. Each group also had a discussion phase at the start of the session, so that players could familiarize themselves with the rules. The video was recorded using three horizontally mounted cameras. Audio was recorded from head-mounted microphones and an array microphone placed at the center at 48kHz. Figures 1 and 2 show a sample of the visual data and the game room layout respectively. The maximum duration of a game-day was set to be $2\times$ (No. of players). This gave us 53.82 hours of audio data. In total, the recording lasted for 3 hours.

For our work, we assigned players to one of the five "Native" classes for every game-day, based on their roles (see Table 1). Four "Derived" classes were defined by combining these native classes, for the purpose of defining deceptive and suspicious behavior (listed in Table 2). The data from *DI* was not used in our experiments.

Fig. 1. Video sample from right, front and left cameras (Right-Left)

3. ESTIMATION TASKS

In this paper, we address two tasks:

Deception and Suspicious Behavior Detection: For this problem, we define two classification problems. The first one is between the liars (*L*) and non-liars (*NL*), which is the standard deception detection task. The second one is between the suspicious behaviour (*SB*) and normal behaviour (*NB*), which is based on the ground-truth obtained from the participants' decisions. This task enables us to find behavior that the players perceive to be suspicious, regardless of whether they are playing deceptive roles or not.

Class	Pts	Description
Successful Liar (SL)	26	A werewolf surviving the game-day without being lynched.
Unsuccessful Liar (USL)	10	A werewolf lynched by the villagers during the game-day.
Suspicious Villager (SV)	11	A villager mistaken for a werewolf and lynched during the game-day
Normal Villager (NV)	105	A villager not killed in the game-day
Dead/Invalid Player (DI)	-	A player killed in one of the previous phases or not present

Table 1. Native Classes

Class	Pts	Description
Liar (L)	36	Werewolves, $\{SL\} \cup \{USL\}$)
Non-Liar (NL)	116	Villagers, $\{SV\} \cup \{NV\}$
Suspicious Behavior (SB)	21	Lynched players, $\{USL\} \cup \{SV\}$
Normal Behavior (NB)	131	Players not lynched, $\{SL\} \cup \{NV\}$

Table 2. Derived Classes

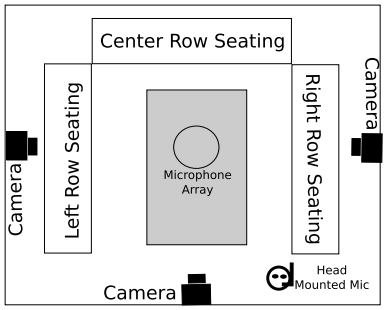


Fig. 2. Recording room setup

Outcome Prediction: In task of predicting the decisions made in every game-day amounts to the detection of players that belong to either USL or the SV classes. This is similar to the suspicious behaviour detection task described previously with one main difference. Here, the choice of suspicious people is narrowed down to the players who are still alive in the given game day. Therefore, a single person is selected from all the alive players, who has the highest chance of being lynched by the players in that game-day.

4. FEATURE EXTRACTION

The audio channels from head-mounted microphones were downsampled to 8kHz and passed through a voice activity detector described in [12]. The output of the voice activity detector was represented in a binary format, based on decisions every 12.5ms. A value of "1" for a frame indicated that the player spoke in that frame. In this context, we define the following:

1. *Speaker Turn* as a continuous duration of a player's speech.
2. A *Successful Interruption* of a speaker i by a speaker j to occur when speaker j starts to speak before a turn of speaker i is over and continues to speak even after that is over.
3. An *Unsuccessful Interruption* of a speaker i by a speaker j to occur when speaker j starts to speak before a turn of speaker i is over and stops speaking before that turn is over.

These concepts are illustrated in Figure 3 and the features we used for our experiments are listed in Table 3. The audio features were extracted once every game-day, for every player (listed in Table 3).

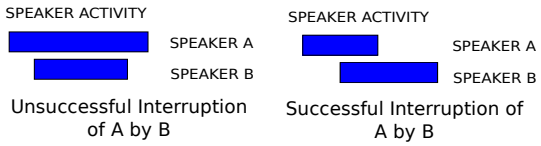


Fig. 3. Illustration of interruptions and turn transfers

Feature Name	Description
Total Speaking Length (TSL)	Ratio of number of frames with speaker-activity to the duration of the day.
Total Speaking Turns (TST)	The number of turns normalized by the duration of the game-day.
Unsuccessful Interruptions Received (UIR)	The number times unsuccessfully interrupted by other players.
Interruption Activity (IA)	Sum of the number of interruptions made and received.
μ, σ of Pitch (mF0, sF0)	Mean and standard deviation (SD) of the median pitch values for each turn (computed using [13]).
μ, σ of Speaking Rate (mSR, sSR)	Mean and SD value of speaking rate (computed from [14])
μ, σ of Power (PR, sPR)	Mean and SD value of the speech signal power [13]
μ, σ of Power Ratio (PRT, sPRT)	Mean and SD value of ratio of aperiodic to periodic power in the speech signal [13]

Table 3. Features extracted per game-day (μ = mean, σ=Standard deviation) .

5. OUR APPROACH

Deception and Suspicious Behavior Detection: For this experiment, we computed the z-score of the features, so that we capture the change in their values from the overall mean, for a given player. The assumption behind this is that the lying or suspicious behavior will correspond to features that deviate a lot from the mean of all the features for that player. When fusing features, PCA based dimensionality reduction was done to reduce complexity and to remove uninformative dimensions. Due to the high imbalance in the number of data points between classes, random sub sampling was used to obtain a balanced training and testing set with equal data points from the native classes. Relevance vector machines [15] with simple Gaussian kernels were used for classification. A leave-4-out cross validation was done, with random sampling and 200 trials.

Outcome Prediction: In order to predict the player lynched at the end of each game day, we analyzed several speaker-turn based features, that are described in section 4. We hypothesized that players who are the most active in earlier rounds of the game have the highest probability of being lynched because players have very little knowledge about the roles of other players early on in the game. Therefore, the outcomes will be strongly dependent on the nature of conversation the players have amongst themselves. Players who are subject to a lot of $interruptions$ are either justifying their position as non-liars, or are involved in the act of questioning or accusing others.

Hence, we ranked players based on their speaker turn features in descending order. We expected to see that one of the players who appear at the top of this list (either in the top 2 or top 3), as described before, will be in the act of justifying him/herself, as a result of which, he/she will be eliminated from the round. We hypothesized that the other top ranking players would tend to be involved more in the interrogation or accusation process. Games after the third day were removed since selecting a subset of two or three players from the four that are remaining is difficult to make any conclusive judgments on. Also, knowledge gained by the villagers with special roles (seer and the little girl) might be used, since they gain significant knowledge in the two or more night phases that precede these days.

6. EXPERIMENTS AND RESULTS

Deception and Suspicious Behavior Detection: The classification performance of the best individual features for both the problems is given in Table 4. Since we were using a balanced training and testing set, the baseline would have a mean F-measure of 0.33, which is obtained by choosing a single class always, giving F-measures for the two classes as 0.67 and 0. When observing the performance of single cues, we found that TSL performed slightly better, than other cues. We found that the performance of the classifier was greatly enhanced by fusing all the features, with the best performance of 0.62 for deception detection. Comparable results are also seen for suspicious behavior detection, where one would expect that for suspicious behaviour detection, people would tend to use more consistent methods to detect who they perceived to be suspicious. We found that when judged on a per-game basis, people were lynched a liar 52.3% of the time. Some of these successful lynchings would have been due to knowledge from the seer or little girl and also cumulative knowledge gained as the game progressed.

To compare our results with findings from social psychology where experiments tended to be carried out in dyadic scenarios, we use the standard mean differences (SMDs), used extensively as an effect-measure in psychology on lie detection [5]. Hence, we computed SMDs for the *L* and *NL* classes using the equation:

$$d = \frac{(\mu_1 - \mu_2)}{V} \text{ Where } V = \sqrt{\frac{(N_1-1)\times\sigma_1^2 + (N_2-1)\times\sigma_2^2}{(N_1+N_2-2)}}$$

Here $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, N_1, N_2$ are the means, variances and the number of data points of the distributions of the two classes that are being compared. We found that the SMDs for the L and NL classes conformed to those found by social psychologists [5], which is interesting since our scenario involves large group discussions. The best SMDs were obtained for $mF0$ (d=0.1) and PR (d=0.204), showing that the distribution of pitch and energy values are higher for liars. Also, TSL and TST exhibited negative values of SMD. This conforms with prior work, as it has been commonly found that liars tend to speak less, in order to avoid attention and making mistakes. We also computed the SMDs for the *NB* and *SB* classes. We found that TST (d=0.591) and TSL (d=0.414) showed the largest effect sizes.

Feature	L vs. NL	SB vs. NB
mF0	0.44	0.38
PR	0.44	0.39
TSL	**0.45**	0.39
PRT	0.44	**0.41**
Fusion	**0.62**	**0.60**
Baseline	0.33	0.33

Table 4. Mean F-measures for the two classes using RVM classifier

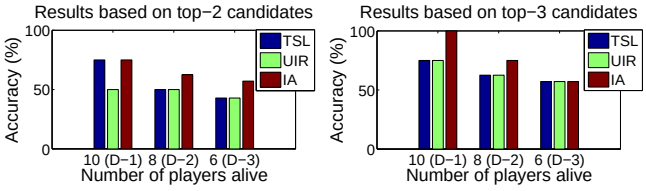
Feature	Top-2	Top-3
TSL	52.63	63.16
UIR	47.36	63.16
IA	**63.16**	**73.68**

Table 5. Average performance for predicting the lynched player.

This shows that the players generally relied on higher speaker activity for lynching in the game-day. This also motivated us to investigate TSL as a feature for outcome prediction. It should be noted that these SMDs were computed for the unbalanced dataset, and does not reflect how good the average classifier performance on the balanced classes can be. We can however infer from the current results that suspicious behaviour might show larger observable changes .

Outcome Prediction: We were able to successfully confirm our hypothesis on the correlation between the activity of the player and the lynchings in the game-day. Players with high *TST*, *IA* and *UIR* were at the highest risk of being killed. The results based on the top-2 and top-3 candidates are shown in Table 5. *IA* gives the best prediction, which also reflects the findings of Raducanu et al. [2].

We analysed the results further by dividing the results depending on the game phase (shown in Figure 4(a)). We see that the accuracy of our predictions improved slightly with the length of the game-day for all features. The prediction for the first two days is high. This result confirms our hypothesis that the decisions will be strongly dependent on speaker-turns in the first few rounds of a game. We also observed similar trends for top-3 candidate (given in Figure 4(b)) with an improvement in performance. This could mean that there is more than one person, who is in the act of accusing others, or, alternatively, the decision is based on the conversation between more than one person with the suspect. Finally, we see that best strategy to remain alive in the initial rounds would be to avoid attention and maintain a low activity, which is characterized by low values in *TST*, *IA* and *UIR*.



(a) (b)

Fig. 4. Accuracy of prediction vs. Progress of game (a) Top-2 ranked (b) Top-3 ranked features.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we have successfully addressed two tasks. Firstly, we have shown the possibility of automatically detecting deception and suspicious behavior, as perceived by participants in social interactions. The preliminary analysis has shown that non-verbal audio cues can be used as indicators of deception and suspicious behavior. Secondly, we have shown that the degree of speaker activity in the initial phase of the game influences the decision of the group. In the last round, however, the decision can be influenced by several factors, such as the information seer might have (if alive), behavior patterns observed by players in the previous rounds etc, which is a subject for further investigation.

In continuation to the present work, we plan on increasing the size of the data set, by recording more games. Additionally, we wish to explore other non-verbal audio-visual features used for conversation analysis. A larger data set would also enable us to statistically model the roles and outcomes in the game. Lastly, this study also is a prelude to a more detailed analysis of behavioral changes in players within and across games, and their effects on decision making.

ACKNOWLEDGMENTS We thank Daniel Gatica-Perez and John Dines for their valuable guidance. Thanks to Bastien Crettol and Olivier Masson for their assistance with the data collection.

8. REFERENCES

[1] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social signal processing: state-of-the-art and future perspectives of an emerging domain," in *Proc. ACM MM*, 2008.

[2] B. Raducanu, J. Vitria, and D. Gatica-Perez, "You are fired! nonverbal role analysis in competitive meetings," *Proc. ICASSP*, April, 2009.

[3] S. Favre, H. Salamin, J. Dines, and A. Vinciarelli, "Role recognition in multiparty recordings using social affiliaation networks and discrete distributions," *Proc. ACM Conf. on Multimodal Interfaces*, pp. 61–68, 2008.

[4] J.E McGrath, *Groups: Interaction and Performance*, Prentice Hall, 1984.

[5] A. Vrij, *Detecting Lies and Deceit: Pitfalls and Opportunities, 2nd Edition*, Wiley, 2nd edition, March 2008.

[6] J. Hirschberg, S. Benus, and et. al, "Distinguishing deceptive and non-deceptive speech," *Proc. Eurospeech*, 2005.

[7] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, "Combining prosodic lexical and cepstral systems for deceptive speech detection," *Proc. ICASSP*, 2005.

[8] T.O Meservy and et.al, "Deception detection through automatic, unobtrusive analysis of nonverbal behaviour," *IEEE Intelligent Systems*.

[9] M. Sung and A. Pentland, "Pokermetrics: Stress and lie detection through non-invasive psysiological sesnsing," Tech. Rep., MIT Media Lab, 2005.

[10] Zarf, "Werewolf party game," `http://www.eblong.com/zarf/werewolf.html`, Accessed August 27, 2009.

[11] M. Braverman, O. Etesami, and E. Mossel, "Mafia: A theoretical study of players and coalitions in a partial information environment," *Annals of Applied Probability*, 2008.

[12] J. Dines, J. Vepa, and T. Hain, "The segmentation of multichannel meeting recordings for automatic speech recognition," *Proc. Interspeech*, 2006.

[13] A. De Chevigne and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *Acoustical Society of America*, 2002.

[14] Nelson Morgon, "mrate estimater," `http://www.icsi.berkeley.edu/ftp/global/pub/speech/morgan/`, Accessed July 25, 2009.

[15] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, August 2006.