The Idiap Wolf Corpus: Exploring Group Behaviour in a Competitive Role-Playing Game

Hayley Hung*
Informatics Institute, University of Amsterdam
Science Park 904
1098 XG , The Netherlands
H.Hung@uva.nl

Gokul Chittaranjan
IDIAP Research Institute
Rue Marconi 19
CH-1920 Switzerland
Gokul.Thattaguppa@idiap.ch

ABSTRACT

In this paper we present the Idiap Wolf Database. This is a audio-visual corpus containing natural conversational data of volunteers who took part in a competitive role-playing game. Four groups of 8-12 people were recorded. In total, just over 7 hours of interactive conversational data was collected. The data has been annotated in terms of the roles and outcomes of the game. There are 371 examples of different roles played over 50 games. Recordings were made with headset microphones, an 8-microphone array, and 3 video cameras and are fully synchronised. The novelty of this data is that some players have deceptive roles and the participants do not know what roles other people play.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing Methods

General Terms

Multimodal, Database

Keywords

deception, human behaviour, multi-party, corpus

1. BACKGROUND

In recent years, the automatic analysis of human behaviour in group conversational settings has become a growing area of research [11, 4, 9, 5, 13, 6, 8, 14]. Different corpora have been used (*e.g.*, the M4 meeting corpus,[1] and the AMI corpus [2]), or created for this purpose (*e.g.*, the Mission Survival Corpus,[3] Free Talk,[4] or the Canal 9 Political Debates[5]) While behavioural traits such as personality [4], dominance [8], and interest [6] have been addressed under multi-party conversational conditions, to our knowledge, there has been no multi-modal corpus that has been used to capture information in a competitive setting where the roles of each of the participants is not universally transparent. The competitive aspect of the game means that some players will need to deceive others which will lead to their behaviour being affected by the stress of having to lie. This corpus serves to enrich the pool of behaviours that can be observed and analysed using existing publicly available corpora.

Investigating deception in a controlled environment is not an entirely new research area. The Columbia/SRI/Colorado (CSC) Corpus [7] was developed for the purposes of providing clean train and test data for analysing deceptive speech. In their scenario, conversations were recorded in dyads where a subject was encouraged to lie to a confederate. In this scenario, the subjects were led to believe that the confederate was unaware of possible deception. Experiments were carried out on the corpus by assuming that deceptive behaviour could be detected by changes in the subject's speech patterns. Other research has been carried out in the social sciences to understand deceptive behaviour but also using dyadic conversational data. To our knowledge, the Idiap Wolf Database is the first dataset recorded that captures deceptive behaviour in a group conversational scenario.

2. THE "WEREWOLF" GAME

The scenario of the data is mainly focussed on a group of people playing a conversational role playing game. "Are you a Werewolf?" is an RPG suited for large groups and is a game of accusations, lying, second-guessing, assassination and mob hysteria [12]. The game is directed by a narrator and the players are randomly divided into villagers and werewolves; some villagers can also have special roles (explained subsequently).

The game proceeds in two alternating phases. (1) The night-phase; in which the villagers are asleep (have their eyes closed) and the werewolves kill a villager of their choice (they make their decision in a discrete manner, so that it is not revealed to anybody except for the narrator). (2) The day-phase; the narrator reveals who the werewolves killed. In this phase, all alive players can talk freely and decide whom they believe to be a werewolf. The werewolves in this phase have to protect their real identity by trying to get an innocent villager lynched, without attracting suspicion. Players who are still alive decide collectively by voting to decide who should be lynched. The voting must be cast before the end of that day. In each day phase, the total time that the players can discuss is calculated by accumulating 2

*The author was working at Idiap Research Institute, Switzerland when this work was carried out.
[1] http://www.idiap.ch/mmm/corpora/m4-corpus
[2] http://corpus.amiproject.org/
[3] http://i3.fbk.eu/en/resources/ms2
[4] http://freetalk-db.sspnet.eu/files/
[5] http://canal9-db.sspnet.eu

MM'10, October 25–29, 2010, Firenze, Italy.

Figure 1: Video sample from right, front and left cameras (Right-Left). For privacy reasons, this is the sole image that can be shown from the data set.

minutes per alive player. If a lynching is not made before the time is up, no-one is killed. Once someone is lynched, they are out of the game and may not speak. The game ends when all werewolves are killed (villagers win) or the number of werewolves and villagers become equal (werewolves win).

Some players in the game can have several special roles [12]. Of these, we used two special roles that help villagers identify the werewolves. (1) The seer - who can determine the true identity of any one person during the night-phase. This is done by the narrator asking the seer to discretely point a person after the werewolves have made their choice in the night-phase. (2) The little girl, who can watch the game with her eyes open in any phase of the game. However, if the little girl's identity becomes known to the werewolves, they almost certainly kill the little girl, to avoid complications. This requires her to be discrete in "spying" other players.

3. DESCRIPTION OF THE DATA

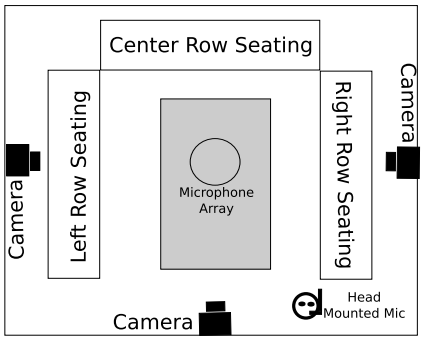

Figure 2: The layout of the recording room

The Idiap Wolf Database can be accessed from `http://www.idiap.ch/dataset/wolf-database`. Our recordings contain 2 werewolves per game, with optionally 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 organization of the corpus is given in Fig. 3. The data set consists of audio-visual recordings of 15 games played by 4 groups of people. The participants were volunteers from the research community at the Idiap Research Institute. In total 36 different people were recorded for the data set. In addition to the players, one person acted as a moderator to guide the proceedings of the game. All participants consented to have their data used for research purposes.

All participants who volunteered for the recordings were not required to prepare anything before playing the game. Each group also had a discussion phase at the start of the session, so that players could familiarize themselves with the rules of the game. Most participants did not know the rules of the game. The rules of the game were written down, and divided into parts that were then randomly distributed

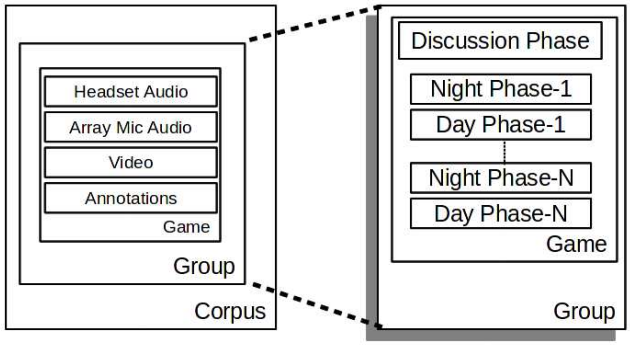

Figure 3: Organization of the Werewolf Corpus

among the participants. Each group spent the initial part of the recording learning the rules of the game together. The rules were distributed randomly to each of the players so that they would need to work together to make sense of the rules (which are provided as a text document in the corpus). This also provided an opportunity to record the group interacting in non-game conditions.

The video was recorded using three horizontally mounted cameras at 30 fps and at a distance from the subjects so that faces are captured at about 30-40 pixels in height. Audio was recorded from head-mounted omni-directional microphones and an array microphone (containing 8 microphones) placed at the centre at 48kHz. Figs. 1 and 2 show a sample of the visual data and the game room layout respectively. A total of 81.17 hours of audio-visual data from 7.3 hours of recordings was collected. In total, there are 50 day-phase games, which are divided per group as shown in Table 1.

Group ID	1	2	3	4
Number of players	10	8	10	12
Number of Day-phases	13	8	10	19

Table 1: The number of day-phases per group.

The average length of each day-phase, depending on the group size or the number of alive players, is shown in Table 2. The summary also shows the number of day phases or data points depending on the size of the group. In general, odd numbers of alive players tended to have fewer data points because there tend to be two-deaths per night-day cycle and games always start with even numbers of players.

4. LABELLING AND ANNOTATION

In the corpus, there are four possible roles that the participants can play in the game. These are shown in Table 3 including the number of examples or data points observed for each role. In addition to these roles, the lynching decision by the group provides labels for each player's role relative to the outcome of each day-phase, as shown in Table 4. Start and end times of each game as well as who is killed in each game phase are also included. Finally, some reference segmentations for speaking activity are provided which is described in more detail in the next section.

5. PRELIMINARY RESULTS

An initial study was carried out on *half* of the data set in [2] using a Relevance Vector Machine [1] that trained a speaker-independent model for classifying the L and NL classes using normalized non-verbal audio cues of all the players in the data set. Using the same method [2], we provide some preliminary results over the complete data. Features were extracted using the audio signal collected from

Number of alive players	3	4	5	6	7	8	9	10	11	12
Number of day-phases	1	3	3	10	4	11	4	8	3	3
Duration (s)	76.00	186.67	67.33	307.7	409.25	437.00	450.75	568.00	418.33	243.33

Table 2: The average day-lengths (seconds) and number of occurrences of different group sizes.

Class	Pts	Description
Liar (L)	81	Werewolves, $\{SL\}\cup\{USL\}$
Non-Liar (NL)	290	Villagers and Special Roles, $\{SV\}\cup\{NV\}$
Special Role 1 (S)	72	Seer, $\{S\}\subset\{NL\}$
Special Role 2 (LG)	36	Little Girl, $\{LG\}\subset\{NL\}$

Table 3: Number of examples (Pts) per role. See Table 4 for other role abbreviations.

Class	Pts	Description
Suspected Player (SP)	42	Lynched players, $\{USL\}\cup\{SV\}$
Normal Player (NP)	329	Players not lynched, $\{SL\}\cup\{NV\}$
Successful Liar (SL)	66	A werewolf surviving the day-phase without being lynched.
Unsuccessful Liar (USL)	15	A werewolf lynched by the villagers during the day-phase.
Suspicious Villager (SV)	27	A villager mistaken for a werewolf and lynched during the day-phase
Normal Villager (NV)	263	A villager not killed in the day-phase

Table 4: Outcome statistics

the headset microphones by sub-sampling the signal to 8kHz at 16-bit precision and passing the result through a voice activity detector (VAD). We used an energy threshold to detect voice activity. The speaking energy of person i in a speech frame f of size K was estimated using the equation:

$$E_{i,f} = \sum_{n=1}^{K} |x_f(n)|, \quad (1)$$

where $x_f(n)$ is the nth sample in frame f and K was chosen as 256 samples (32ms) for 8kHz audio with an overlap of 16ms between consecutive frames.

The energy of all the head-set audio was computed to give a matrix $E \in \mathbb{R}^{N\times F}$, where N is the number of participants in the recording and F is the number of frames. Since the audio channels were synchronized, the following matrix $D \in \mathbb{R}^{N\times F}$ was computed, so that the values of each frame in D containing voice activity from the person wearing the microphone would have a higher value than the frames that do not. For each person i, D_i was defined as,

$$D_i = \sum_{n=1, n\neq i}^{N} E_i - E_n, \quad (2)$$

Feature Name	Description
Total Speaking Length (TSL)	Ratio of number of frames with speaker-activity to the duration of the day.
mean Pitch ($\mu F0$)	Mean and standard deviation (SD) of the median pitch values for each turn (computed using [3]).
mean Speaking Rate (μSR)	Mean and SD value of speaking rate (computed from [10])
Total Energy (TE)	The sum of the energies of all frames on audio in a given speaker turn (see Equation 1).

Table 5: Features extracted per day-phase

where D_i, E_i refer to the i^{th} row of D and E respectively. A threshold was determined empirically to obtain the voice activity in the group using D. The channel having the highest energy for the frames with voice activity was labelled as having voice activity. The output of the voice activity detector is stored in binary format where a "1" represents the presence of voice activity in a 32ms window of the energy. Each window was also shifted by 16ms.

For simplicity, we have assumed that two people are not speaking simultaneously. Previously [2], we used a more sophisticated VAD which could detect multiple simultaneous speakers. We found that there was not a significant difference in both the VAD performance for the features we used and also the liar/ non-liar detection performance.

The VAD was qualitatively satisfactory upon inspection of several random files. Furthermore, we evaluated the output of the VAD using 1-minute of manually segmented audio from an 8-speaker game, during a period when the players were very active in their discussions. The false-rejection rate averaged over all players, for silence and speech was 1.03% and 1.07% respectively. The average, highest and lowest percentage of speech activity in the 1 minute duration was 14.95%, 29% and 1.94% respectively. The reference segmentations are also provided in the corpus.

The output of the voice activity detector and the sub-sampled audio were used to extract a selection of non-verbal features, which are summarised in Table 5. For each feature, the classification experiments were run 200 times on different subsets of the data. The evaluation was carried out in a similar manner to leave-4-out cross validation. However, since there is a large imbalance between the L and NL classes, the training data was under-sampled for the NL class such that the two classes had equal numbers of points. We see that the best performance was achieved when the total energy feature is used. Therefore, for our data, liars tend to have more discriminative differences in their speaking energy. This performed significantly better than the baseline, which would always classify a test example as a liar. Table 6 shows the mean f-measure for 200 runs.

Feature	F_L	F_{NL}	Mean F
TSL	0.63	0.23	0.43
TE	**0.68**	**0.5**	**0.59**
μSR	0.65	0.15	0.4
$\mu F0$	0.63	0.28	0.45
Baseline	0.67	0	0.33

Table 6: Results of automatically detecting liars (L) and non-liars (NL) using RVMs. Results are reported in terms of mean f-measure. See Table 5 for a summary of the feature descriptions.

6. DISCUSSION

Our preliminary tests have only been carried out on two of the possible annotated classes in the data. Tables 3 and 4 show that there are a number of roles and also derived classes based on the outcomes of each day-phase. Therefore, there is scope to carry out experiments on group behaviour such as predicting who will be killed at the end of a day, who will be suspected incorrectly, who is good or bad at lying etc. In addition to studying deceptive behaviour and roles, the data has many more potential applications.

From a behavioural perspective, games were often driven by one or two players who led the collective decision process, providing many examples of agreement and disagreement during the conversations as well as affiliative and confrontational behaviours. As well as persuasive and leadership behaviours, submissive behaviour was also seen from players who were accused of being a wolf. Also, varying levels of excitement were seen over each or progressing day-phases.

While the preliminary tests were carried using just the audio data, the corpus provides an opportunity for the study of audio-visual deceptive behaviour. We believe that this constitutes the largest publicly available data set that could be used to analyse deceptive roles, both in the quantity of data, and the number of participants.

In terms of video feature extraction, some initial tracking tests suggest that a person's face can be tracked but quantitative evaluations are yet to be carried out. All the participants are filmed while seated facing the camera but some leaning can cause more challenging periods where the faces may be concealed or viewed from extreme tilt or pan angles. Given the resolution of each face, it is unlikely that detailed facial expression estimation can be used. Also, there are many examples of pointing that are used in the game during voting for who to lynch or who the wolves want to kill.

For audio feature extraction, all participants were recorded using both head-set microphones and also an 8-microphone array. With the inclusion of a some scripted speech, the data provides good examples of controlled speech, and free speech in both excited and also more calm stages of the conversations using the same speakers.

Aside from the practicalities of the data, the aim was to produce a corpus that captures natural interactive behaviour in a face-to-face scenario. Interviews with participants afterwards indicated that players found the game enjoyable and engaging. The competitive aspect of the game and the participants' willingness to play helped to make their behaviour natural and emotions genuine.

7. CONCLUSION

The Idiap Wolf Database provides a rich corpus for the analysis of natural interactive behaviour in a play scenario. It can also be used for testing feature extraction systems. Initial experiments on the data show that it captures natural behaviour well and can be used for a wider variety of experiments. To our knowledge, this is the only publicly available corpus that provides audio-visual recordings of deceptive behaviour in a group conversation scenario.

Acknowledgements

The collection of this corpus has been supported by European IST Program Project FP6-033812 (AMIDA). Thanks to Bastien Crettol and Olivier Masson for their assistance with the data collection and to Daniel Gatica-Perez for his valuable guidance. We also thank all the volunteers who gave up their free time to take part in these recordings.

8. REFERENCES

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

[2] G. Chittaranjan and H. Hung. Are you a werewolf? detecting deceptive roles and outcomes in a conversational role-playing game. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2010.

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

[4] W. Dong, B. Lepri, A. Cappelletti, A. S. Pentland, F. Pianesi, and M. Zancanaro. Using the influence model to recognize functional roles in meetings. In *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*, pages 271–278, New York, NY, USA, 2007. ACM.

[5] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image Vision Comput.*, 27(12):1775–1787, 2009.

[6] 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 (ICASSP)*, Philadelphia, Mar. 2005.

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

[8] D. Jayagopi, H. Hung, C. Yeo, and D. GaticaPerez. Modeling dominance in group conversations from non-verbal activity cues. *Special issue on Multimedia in IEEE Trans. on Audio, Speech and Language Processing*.

[9] N. Mana, B. Lepri, P. Chippendale, A. Cappelletti, F. Pianesi, P. Svaizer, and M. Zancanaro. Multimodal corpus of multi-party meetings for automatic social behavior analysis and personality traits detection. In *TMR '07: Proceedings of the 2007 workshop on Tagging, mining and retrieval of human related activity information*, pages 9–14, New York, NY, USA, 2007. ACM.

[10] N. Morgon. mrate estimater. `http://www.icsi.berkeley.edu/ftp/global/pub/speech/morgan/`, Accessed 25/07/2009.

[11] K. Otsuka, H. Sawada, and J. Yamato. Automatic inference of cross-modal nonverbal interactions in multiparty conversations: "who responds to whom, when, and how?" from gaze, head gestures, and utterances. In *International conference on Multimodal Interfaces*, pages 255–262, New York, NY, USA, 2007. ACM.

[12] A. Plotkin. Werewolf party game. `http://www.eblong.com/zarf/werewolf.html`, Accessed August 27, 2009.

[13] H. Salamin, S. Favre, and A. Vinciarelli. Automatic role recognition in multiparty recordings: using social affiliation networks for feature extraction. *IEEE Transactions on Multimedia*, 11(7):1373–1380, 2009.

[14] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, December 2008.