LieCatcher: Game Framework for Collecting Human Judgments of Deceptive Speech

Sarah Ita Levitan, James Shin, Ivy Chen, Julia Hirschberg
Dept. of Computer Science, Columbia University
New York, NY USA
{sarahita@cs.columbia.edu, js4785@columbia.edu, ic2389@columbia.edu, julia@cs.columbia.edu}

Abstract

We introduce ”LieCatcher”, a single-player web-based Game With A Purpose (GWAP) that allows players to assess their lie detection skills, while simultaneously providing human judgments of deceptive speech. Players listen to audio recordings from the Columbia X-Cultural Deception (CXD) Corpus, a collection of deceptive and non-deceptive interview dialogues, and guess if the speaker is lying or telling the truth. They are awarded points for correct guesses, and lose lives for incorrect guesses, and at the end of the game, receive a score report summarizing their performance at lie detection. We present the game design and implementation, and discuss plans for using the human annotations for research into the acoustic-prosodic properties of believable, trustworthy speech. This game framework is flexible and can be applied to other useful speech annotation tasks, and we plan to make the game available to the public to extend for other tasks.

Keywords: games with a purpose, speech annotation, deception, trust

1. Introduction

In recent years, much progress has been made in developing and improving human language technologies. Some of these advances have been made using supervised learning methods, which rely on an abundance of annotated data. For example, a state-of-the-art commercial automatic speech recognition (ASR) system can rely on as much as 5000 hours of annotated speech (Hannun et al., 2014) Speech corpus annotation is a critical component of any speech related research. Traditionally, this annotation has been done by a small group of highly skilled annotators. This is a time consuming process, with extensive training required, and it is also expensive. In recent years, crowdsourcing has revolutionized the annotation process. Instead of relying on a few skilled annotators, crowdsourcing allows us to collect annotations from a large group of unskilled crowd workers, quickly and cheaply. Because this work is unskilled, it is important to take steps to control the quality of the annotations. An alternative approach to collecting annotations involves the use of Games With A Purpose (GWAP). The idea behind GWAP is to motivate people to solve computational problems by presenting the problem as a series of simple steps in an enjoyable game format.

In this work we have designed and implemented a GWAP with the goal of collecting human judgments for a corpus of deceptive speech. In our ongoing research, we are examining human ability at deception detection. The corpus contains dialogues between interviewer/interviewee pairs, where the interviewer asks 24 biographical questions, and the interviewee aims to deceive her partner for a random half of the questions. The interviewer records his judgment of each question, i.e. whether he thinks his partner is telling a lie or the truth. With this paradigm, we have record of a single human judge for every interviewee response. However, we are interested in exploring human perception of deception at a larger scale, exploring individual differences in how people perceive deception, as well as exploring trust. To do this, we need many instances of human judgments for each utterance. A previous perception study of human performance at deception detection recruited 32 participants to listen to audio recordings ranging from 25-50 minutes long, and annotate them with their judgments of deception (Enos et al., 2006). This process typically requires an experimenter to schedule, train, and supervise the participants, and it can be a time consuming and expensive ordeal. In addition, although the human judges are paid for their time, there is no explicit motivation for the judges to perform well at the specific task that they are working on, and it is conceivable that they will become disinterested in the task and even answer randomly.

Here we introduce a GWAP to collect large-scale human annotations of deception. This framework has several advantages. It enables the rapid, large-scale collection of human annotations - multiple users can play in parallel, and they can play the game from any location, at any time. It is inexpensive - players are unpaid, motivated by the enjoyment of the game, and there is no need for a human to train the players. There is explicit incentive for players to perform well at the task, in the form of points and loss of game lives. In addition, the game implementation is flexible and makes it easy to manipulate conditions, so that we can design experiments to test theories of human perception of deception.

The rest of this paper is organized as follows. Section 2. reviews related work, and Section 3. details the speech corpus that we use for the game. In Section 4., we describe the design and implementation of LieCatcher. Section 5. describes an initial pilot study that we conducted to get early feedback about the game design. We conclude in Section 6. with a discussion of ongoing and future work.

2. Related Work

Games with a purpose (GWAP) have been previously used for annotation of language corpora, including text and speech modalities. ”tashkeelWAP” (Kassem et al., 2016) is a web application with a single-player and a two-player game where Arabic speaking players digitize Arabic words with their diacritics that were not correctly recognized by OCR systems. ”Phrase Detectives” [1] is another annotation game, where players label relationships between words and phrases, to create a rich language resource of anaphoric co-references (Chamberlain et al., 2008).

”Voice Race” (McGraw et al., 2009) and ”Voice Scatter” (Gruenstein et al., 2009) are GWAP that are educational for their players, and also useful for obtaining speech annotations. In ”Voice Race”, A player is presented with a set of word definitions on flashcards, and they must quickly say the corresponding words. In ”Voice Scatter”, the player chooses flashcards to study, and when presented with a term, speaks the definition into a microphone, earning points for correct responses. This game elicits spontaneous speech in longer sentences. By using speech recognition as well as contextual information from the games, the spoken utterances can be labeled orthographically with near perfect accuracy. These games are enjoyable as well as educational, and provide labeled speech data as a by-product of the games.

3. Corpus

For this work, we examined the Columbia X-Cultural Deception (CXD) Corpus (Levitan et al., 2015) a collection of within-subject deceptive and non-deceptive speech from native speakers of Standard American English (SAE) and Mandarin Chinese (MC), all speaking in English. The corpus contains dialogues between 340 subjects. A variation of a fake resume paradigm was used to collect the data. Previously unacquainted pairs of subjects played a ”lying game” with each other. Each subject filled out a 24-item biographical questionnaire and were instructed to create false answers for a random half of the questions. They also reported demographic information including gender and native language, and completed the NEO-FFI personality inventory (Costa and McCrae, 1989).

The lying game was recorded in a sound booth. For the first half of the game, one subject assumed the role of the interviewer, while the other answered the biographical questions, lying for half and telling the truth for the other; questions chosen in each category were balanced across the corpus. For the second half of the game, the subjects' roles were reversed, and the interviewer became the interviewee. During the game, the interviewer was allowed to ask the 24 questions in any order s/he chose; the interviewer was also encouraged to ask follow-up questions to aid them in determining the truth of the interviewee's answers. Interviewers recorded their judgments for each of the 24 questions, providing information about human perception of deception. The entire corpus was orthographically transcribed using the Amazon Mechanical Turk (AMT)[2] crowd-sourcing platform, and the speech was segmented into *inter-pausal units* (IPUs), defined as pause-free segments of speech separated by a minimum pause length of 50 ms. The speech was also segmented into turn units, where a turn is defined as a maximal sequence of IPUs from a single speaker without any interlocutor speech that is not a *backchannel*. Finally, the speech was segmented into question/answer pairs, using a question detection and identification system (Maredia et al., 2017) that employs word embeddings to match semantically similar variations of questions to a target question list. This was necessary because interviewers asked the 24 questions using different wording from the original list of questions.

In total, there are 7,141 question/answer pairs, each associated with a question number (1-24), start and end times in the full session recording, transcribed text, and truth value (T or F).

4. Game Design and Implementation

4.1. Game Design

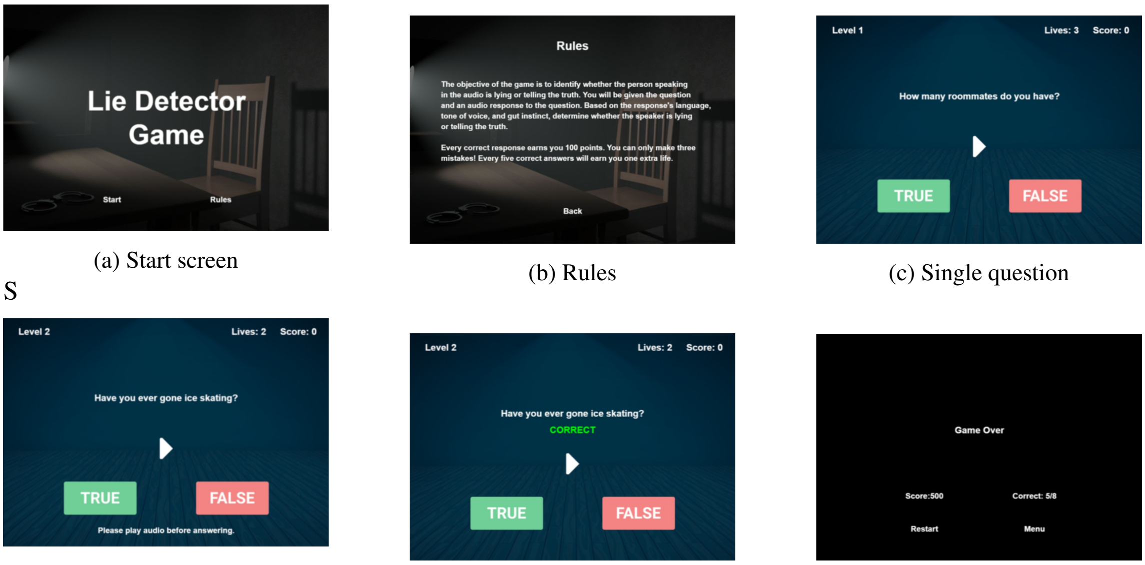
The game design is simple and flexible. The player is presented with a series of audio recordings from the CXD corpus, each one paired with the text of the interviewer question that prompted the interviewee's response. The player listens to each interviewee audio clip, and selects whether they think the speaker is lying or telling the truth. The player can listen to the audio an unlimited number of times, but is required to listen to the full audio before selecting a ”True” or ”False” button. Each player is given 3 lives; a correct guess earns the player 100 points, while an incorrect judgment causes the player to lose one life. The game ends when the player has lost 3 lives, and the final screen of the game is a display summarizing the player's performance. The points and lives, as well as the final score summary, serve to motivate the player to try their best to succeed at the game.

Figure 1 displays screenshots from the main 6 stages of the game: (a) Start screen, where users select ”play” or ”rules”, (b) Rules, which lists the rules of the game, (c) Single question, which shows the text of a question along with a play button to listen to the audio, along with ”True” and ”False” buttons to select the deception judgment, (d) Error message displayed if the audio was not played before selecting a button, (e) Feedback after the player selects a button, showing the correct answer, and (f) Game over and score report displaying information about player performance when the player loses all his lives.

There are many decisions to make in creating this framework. How many lives should players start with? How many times can the player listen to the audio? Should the players receive instant feedback about their judgments, or only at the end of the game? Should the audio clips be randomly chose, or perhaps ranked in some manner (e.g. by difficulty)? Some of these decisions may significantly impact player performance. For example, it is possible that

[1] http://anawiki.essex.ac.uk/phrasedetectives/

[2] https://www.mturk.com/mturk/



The objective of the game is to identify whether the person speaking in the audio is lying or telling the truth. You will be given the question and an audio response to the question. Based on the response's language, tone of voice, and gut instinct, determine whether the speaker is lying or telling the truth.

Every correct response earns you 100 points. You can only make three mistakes! Every five correct answers will earn you one extra life.

How many roommates do you have?

Have you ever gone ice skating?

Please play audio before answering.

(a) Start screen

S

(b) Rules

(c) Single question

(d) Error message if audio not played

(e) Feedback after response

(f) Game over and score report

Figure 1: Main states of the game.

players would benefit from receiving instant feedback about their judgments as they play the game. We are interested in exploring the effect of these parameters, and therefore have implemented these options in a flexible manner, so that we can experiment with different settings and observe their effects. In addition, we plan to extend the game to accommodate for different levels and designs (where certain levels would have differing conditions – limited time, or limited number of listens of the audio, or no feedback, or more difficult audio recordings), which should be interesting to study as the data grows and the game is played more frequently over time.

4.2. Game Implementation

We used PhaserJS [3] for LieCatcher's framework. PhaserJS is a 2D game framework for creating HTML5 games for web browsers. We chose this framework because of its lightweight features and intuitive javascript syntax. PhaserJS is a state-based game framework, meant to support small games. In a state-based game framework, every scene in the game is its own state that the user is in (i.e., ”Menu”, ”Rules”, ”Stage1”, ”Stage2”, etc.). Because of this, assets must be loaded quickly, so as not to slow the gameplay. Other larger game development engines as Unity allow support for multithreaded applications, but this comes with additional overhead, and is not necessary for our lightweight game.

For the backend of the game, we stored the audio files in a MongoDB database [4] hosted on our own server. Since PhaserJS does not natively support database queries, we set up endpoints on our site server using ExpressJS URLs that returned queries from our database. When loading assets, phaserJS queries the appropriate site server endpoints and receives request responses corresponding to the data of interest. Specifically, in each state, loading assets is typically done synchronously in a ”pre-load” method before the assets are placed into the scene. Because it takes a significant amount of time to load over 7,000 audio files, we instead loaded the audio files asynchronously in a queue in the background during gameplay, as to not interfere with the user experience. One audio file is loaded in the background while the player plays each stage (i.e. each audio clip). The weakness to this approach is that a player may spend less time playing a certain stage than the time it takes to load one audio file. However, the longest audio files load in under 5 seconds, so loading times are not a major issue to the user experience.

When a player loses all three lives, they are sent to the game over screen, and during this state, we send user session data to the user database. The data include the IDs of questions that were correct and incorrect, the time it took to answer each question, the player score, the date, and the number correct and total answered. We used the JS fetch API as a request handler to pass JSON data into the request bodies. This was done to collect data from the user session and store it into a separate user database.

5. Pilot Study

In order to get early feedback about the game, we conducted a pilot study where 40 students played the game and answered a pre-game and post-game survey. For the purpose of the study, we structured the game as 2 levels, with 10 audio samples in each level. In level 1, players were not provided with any feedback about the correctness of deception judgments. That is, they received no points for correct judgments, did not lose lives for incorrect judgments, and there was no message on the screen to indicate whether their judgment was correct or incorrect. In level 2, players received immediate feedback about their judgments with a displayed ”correct” or ”incorrect” message, as well as earning 100 points for each correct judgment. At the end of the 10 audio clips in level 2, players were given a score

[3] https://phaser.io

[4] https://www.mongodb.com

report for the level 2 questions.

Before playing the game, players filled out a pre-game survey. They were asked to report their gender and first language spoken, and answered three questions: (1) How often can you spot a lie in daily life? (on a scale of 1 to 5, with 1 being almost never and 5 being almost always) (2) How often do you think people lie in daily life in order to achieve some gain, either material or social? (also on a scale of 1 to 5) (3) Do you have experience in law enforcement or in another job where spotting lies is important? If yes, please describe.

After answering the pre-game survey, participants played the pilot game. We introduced two quality control questions to ensure that players were paying attention and listening to the audio, and not selecting buttons randomly (e.g. with their audio turned off). In each level, one of the audio segments was a recording that said ”Please wait 5 seconds and select TRUE” or ”select FALSE”.

After playing the game, participants answered a post-game survey and provided feedback about their experience. Questions included: Did you find the game to be easy to use? Which level did you prefer (level 1 or level 2)? How would you rate your ability to detect deception after playing this game? How well do you think your score on the game reflects your ability to detect lies in the real world? Did you like the premise of the game? Would you recommend the game to a friend? Did you like the game graphics? Players also provided feedback about the quality control questions, and general ideas about the game. They also reported strategies that they used in making their judgments. Some of the survey questions were adapted from a study of human judgments of deception by Enos et al (Enos et al., 2006) and others game evaluation questions were adapted from (Sturm et al., 2017).

5.1. Pilot Study Survey Responses

40 students participated in the pilot study, 26 female and 14 male. 77% of the participants were native speakers of English, and the rest were native speakers of other languages (e.g. Chinese), but were proficient in English. Only one player reported job experience with lie detection.

35 of the players reported using a laptop or desktop computer to play, while 5 players used a mobile device. They played using various browsers, including Chrome, Firefox and Safari, without compatibility issues. Overall, the feedback about the game was positive. 85% found the game easy to use, and 75% reported that they would or might recommend the game to a friend.

Player responses were mixed about whether they thought the game is a good way to assess ability to detect lies. 57% responded yes or maybe, while 43% responded no. 73% of players preferred level 2, where feedback was given, to level 1. This information is useful for future game design choices. The feedback about the quality control questions was informative - some players thought it was a great idea to check attention, while others found it slightly confusing. In the future, we might inform players to expect such questions distributed throughout the game, to avoid confusion. 70% of respondents liked the premise of the game, 18% were neutral, and 12% did not like the premise. 50% liked the game graphics, while 35% were neutral and 15% did not like them. Going forward, we plan to incorporate ideas from this initial player feedback in order to improve the player experience.

5.2. Pilot Study Player Behavior

Players were overall 49.86% accurate in their predictions, not including check questions. The minimum correct number of questions by a player was 5 correct, while the maximum was 13. The median and mean was 9 correct, with a standard deviation of 1.94. 100% of players answered the check questions correctly and made sure to listen to directions and wait five seconds, indicating that players were attentive in making their decisions. Overall, however, players were still on average approximately as accurate as random guessing.

There was a noticeable difference in player performance in between the levels. For level 1, the average number of correct questions was 4.1 out of 9, with a median of 4 and standard deviation of 1.18. The overall accuracy of all players was 45%. In contrast, level 2 players averaged 4.9 correct answers of 9, with a median of 5 and standard deviation of 1.27. The overall accuracy for level 2 was 55%.

Some questions had collective responses strongly in favor of an answer choice. In particular, question 5 had 33 responses to ”T” and 7 to ”F” with an accuracy of 17.5%, question 9 had 34 responses to ”T” and 6 to ”F” with an accuracy of 85%, and question 14 had 33 responses to ”T” and 7 to ”F” with an accuracy of 82.5%. There were no questions with responses strongly in favor of ”F”, indicating that for the given audio sample pool, players were more inclined to trust confidently than to accuse.

There was a negligible difference in performance between female and male players. Female players were 50% accurate with a trust rate of 60%, while male players were 49% accurate with a trust rate of 62%.

6. Conclusions and Future Work

We presented LieCatcher, a GWAP where players can learn how well they perform at deception detection, while providing human annotations of deception. This game framework allows for the rapid and large-scale collection of human annotations of deceptive speech, and can easily be extended to other speech annotation tasks. We plan to make the game implementation publicly available for further development. We conducted a pilot study to get early player feedback about the game. The initial feedback is promising, and we plan to incorporate some of the feedback to further improve the game.

We are now in the process of testing the game on student volunteers. So far we have received feedback that the game is entertaining; people enjoy assessing their abilities at lie detection. Once this is completed and preliminary feedback is addressed, we plan to distribute the game on crowdsourcing platforms such as Amazon Mechanical Turk to collect large-scale annotations. After this data collection phase, we will conduct an analysis of acoustic-prosodic properties of trustworthy speech. We also plan to explore the role of gender and culture (of the speaker as well as the listener) on trust.

7. Acknowledgements

This work was partially funded by AFOSR FA9550-11-1-0120 and by NSF DGE-11-44155.

8. Bibliographical References

Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008). Phrase detectives: A web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics '08)*, pages 42–49.

Costa, P. and McCrae, R. (1989). Neo five-factor inventory (neo-ffi). *Odessa, FL: Psychological Assessment Resources*.

Enos, F., Benus, S., Cautin, R. L., Graciarena, M., Hirschberg, J., and Shriberg, E. (2006). Personality factors in human deception detection: comparing human to machine performance. In *INTERSPEECH*.

Gruenstein, A., McGraw, I., and Sutherland, A. (2009). A self-transcribing speech corpus: collecting continuous speech with an online educational game. In *International Workshop on Speech and Language Technology in Education*.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Kassem, L., Sabty, C., Sharaf, N., Bakry, M., and Abdennadher, S. (2016). tashkeelwap: A game with a purpose for digitizing arabic diacritics.

Levitan, S. I., An, G., Wang, M., Mendels, G., Hirschberg, J., Levine, M., and Rosenberg, A. (2015). Cross-cultural production and detection of deception from speech. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pages 1–8. ACM.

Maredia, A. S., Schechtman, K., Levitan, S. I., and Hirschberg, J. (2017). Comparing approaches for automatic question identification. SEM.

McGraw, I., Gruenstein, A., and Sutherland, A. (2009). A self-labeling speech corpus: Collecting spoken words with an online educational game. In *Tenth Annual Conference of the International Speech Communication Association*.

Sturm, D., Zomick, J., Loch, I., and McCloskey, D. (2017). "free will": A serious game to study the organization of the human brain. In *International Conference on Human-Computer Interaction*, pages 178–183. Springer.