*INTERSPEECH 2021*
30 August – 3 September, 2021, Brno, Czechia

# An Agent for Competing with Humans in a Deceptive Game Based on Vocal Cues

*Noa Mansbach*[*]*, Evgeny Hershkovitch Neiterman*[*]*, Amos Azaria*

Computer Science Department, Ariel University, Israel
noa.aizer@msmail.ariel.ac.il

## Abstract

In this work we present the development of an autonomous agent capable of competing with humans in a deception-based game. The agent predicts whether a given statement is true or false based on vocal cues. To this end, we develop a game for collecting a large scale and high quality labeled sound data-set in a controlled environment in English and Hebrew. We develop a model that can detect deception based on vocal statements from the participants of the experiment, and show that the model is more accurate than humans.

We develop an agent that uses the developed deception model and interacts with humans within our deceptive environment. We show that our agent significantly outperforms a simple agent that does not use the deception model; that is, it wins significantly more games when played against human players. In addition, we use our model to detect whether a statement will be perceived as a lie or not by human subjects, based on its vocal cues.

## 1. Introduction

It is hard to overestimate the damage and harm caused by deception and fraud. Indeed deception has caused the loss of lives and property [1, 2]. Not surprisingly, fraud and deception are not limited to a specific culture, and are common worldwide [3]. It is well known that humans are not very good at detecting whether a statement is truthful [4, 5]. Therefore, throughout history people have tried to develop methods for lie detection. In the far history, many cruel methods were used to detect liars (see [6] for several examples of such methods). In 1921 John Augustus Larson invented the polygraph [7], a device intended to detect a lie by recording several body measures, such as breathing rate, pulse, blood pressure, and perspiration. It is assumed that all these measures accelerate while telling a lie. However, these devices require the suspect to be attached to different appliances and cannot be performed retrospectively, or when the suspect is not present.

Autonomous agents interacting with humans become more and more ubiquitous, appearing in many different domains, including smart home assistants (such as Alexa, Siri and Google Home), social networks bots (such as Facebook and Twitter [8]), and marketing chat-bots [9]. We believe future intelligent agents must be able to interact in an environment in which humans do not always tell the truth. Similar to the input provided by a human user to smart home assistants, and a question answering environment, we focus on situations in which short sentences are vocally said by humans, and attempt to build an agent for one such environment.

Interacting with humans in a deceptive environment requires human modeling. That is, the development of a model that, given a short sentence, can predict whether the sentence is true or false. Therefore, we attempt to build a human model based on machine learning techniques. However, machine learning methods require a data-set. While there exist several labeled data-sets for deception, due to the nature of the deception detection problem, most of them suffer from one or more of the following: (i) an entire conversation is either marked as true or false, where parts may be true and parts may be false (ii) subjects may be told to lie, and thus their tone of voice may be different than people telling authentic lies (iii) labels may be inconclusive; that is, some sentences may be seen as a true statement by some people and false by others.

We therefore develop a speech based game for collecting a large scale and high quality labeled data-set in controlled environments in English and Hebrew. The models are available at: `https://github.com/NoaAizer/Cheat-Game-Detector.git`, and the data-set is available at: `http://www.azariaa.com/misc/Noa/Interspeech21data.zip`. Using this data-set we develop a model that can detect deception based on vocal statements from the participants, and show that it outperforms humans in detecting deceptive speech. Furthermore, we develop an agent that interacts with humans in the deception game environment. We show that an agent using the deception model significantly outperforms an agent that does not use the deception model; that is, it wins significantly more games when playing against human players. This result indicates that the deception model is reliable enough to be used in practice. In addition, we develop a model that, given a vocal statement, determines whether it is *perceived* as deceptive by humans.

To summarize, our main contributions of this paper are: (i) The gathering of a high-quality multilingual data-set for deception detection. (ii) The development of a deception detection model based on verbal cues, which outperforms humans in terms of accuracy. (iii) The development of an autonomous agent that interacts with humans. This agent uses the model developed in the previous phase. We show that an autonomous agent using our deception detection model significantly outperforms an agent that does not use this model.

## 2. Related Work

Deception detection is a critical problem studied by psychologists, criminologists, and computer scientists [10]. In recent years deception detection has aroused interest in the natural language processing as well as human computer interaction communities. Current models can adopt language and behavior cues, and research examining them to deception has been quite promising. However, the overall detection performance is still not satisfactory, especially when the model is based on speech alone.

Past research in the detection of deception can be broadly

[*] equal contribution



http://dx.doi.org/10.21437/Interspeech.2021-83

classified as Verbal and Non-verbal. In verbal deception detection, the features are based on the linguistic characteristics, such as n-grams and sentence count statistics [11], of the statement by the subject under consideration. Mihalcea and Pulman [11] show that the use of Linguistic Inquiry and Word Count (LIWC) lexicon are helpful in detecting deceptive behavior. There exist several data-sets for deception detection in text. Ruiter and Kachergis [12] collected data from the Mafiascum website. The data-set consists of chat correspondence within a game called "Mafia". In this game there is a public chat open to all participants and a private chat open only to a small group called "The Mafia". The members of The Mafia attempt to hide their identity while the rest for the players goal is to expose them. Perez and Mihalcea [13] composed a data-set using Mechanical Turk. Paid users were asked to supply seven true statements and seven lie statements.

In non-verbal deception detection, physiological measures were the main source of signals for detecting deceptive behavior. Polygraph tests measure physiological features such as heart rate, respiration rate and skin temperature of the subject under investigation [14]. However, administering these tests requires physical presence and the use of intrusive equipment. Other non-verbal deception detection methods include video and voice. However, very few studies have attempted to detect deception based on voice alone.

Several works have used data-sets for deception detection that include voice [15, 16, 17]. However, all these data-sets are based on conversation and storytelling scenarios. Nasri et al. [15] asked their subjects to tell two stories in a recording studio, while trying to confuse the listener with which story is true. The Colombia SRI Colorado (CSC) corpus [17], was collected using one on one interviews. The subjects' goal was to convince the interviewer that they fit a given profile. The interviewer asked the subjects follow-up questions in order to investigate if some statements are truthful or not. For each sentence, the subjects were asked to indicate whether the reply was true or contained any false information. We note that a story labeled as a lie may contain some details that are true while others are false, thus making the data-set less accurate. In this paper, we present a data-set consisting of short speech utterances with ground truth labels, thus, making it more robust and applicable for short question answering.

Krishnamurthy et al. [18] introduce a deep learning approach for deception detection based on multimodal inputs, i.e., video, audio and text. They use the real-life trial data-set collected by Pérez-Rosas et al. [19]. Their audio model first extracts features by using openSMILE, which outputs 6373 features for every audio; these features serve as an input to a multilayered perception with 300 neurons in the hidden layer.

As far as human-agent interaction in deceptive environments, we are aware of only two works that tackle the issue [20] and [21], both using text only inputs. These works use variations of the Mafia game described above. To the best of our knowledge we are the only work developing an agent interacting with humans by sound in a deceptive environment.

# 3. Data collection

## 3.1. Collection environment

In this paper we focus on detecting deception in the "cheat game" environment. The "cheat game" (also known as B.S. and the bluff game) is a turn taking card game where the players' goal is to play all of their cards. After dealing eight cards


Figure 1: *The graphical interface developed for the "cheat game".*

to each player, the game begins with a card flipped over from the deck of cards to a pile of cards. On each turn a player may place up-to four cards on the pile of cards; these cards may either contain cards that are one higher than the current card or one lower. The cards placed on the pile are faced down, therefore, the player may claim to put cards that are different from what they actually placed. If a player suspects that their opponent is cheating (i.e. placed cards that are different from what they claimed), the player may call out a cheat. In this situation, if the opponent did actually cheat, this player collects all the cards, otherwise, the player that called out a cheat collects the cards. Instead of placing cards on the pile, a player may draw three cards from the deck[1]. See Figure 1 for a screen-shot of the game.

We recruited two types of subjects: US and Israeli subjects. US subjects were recruited using Amazon's Mechanical Turk service and played the game in English. In addition, we gathered graduate students and staff from a computer Science department in Israel. The Israeli subjects played the game in Hebrew. Each player played three games; every game ended either when one of the players played all her cards or when a 12 minute countdown clock reached zero.

## 3.2. Data-set structure

The data-set was collected from 34 test subjects. 18 played in the English language and 16 played in Hebrew. Additional statistics on the subjects can be found in Table 1.

In total, the data collection phase provided 950 labeled samples. 423 statements were in Hebrew and 527 in English. 598 were true statements and 352 statements were false. We note that the ratio of the false statements (37%) is more balanced than the ratio in other common deception data-sets [22, 12].

# 4. Models

We used a Voice Activity Detector (VAD) to trim the recording of silence and background noise for all samples. Many samples have silence periods at the beginning or end since the game records a fixed duration, and some subjects started speaking af-

[1]In the original game rules a player draws only a single card, however, in order to encourage people to cheat, we raised the number of cards the player must draw to three.

Table 1: *Test subjects details*

| | | |
|---|---|---|
| **Gender** | Male | 18 |
| | Female | 16 |
| **Language** | English (live in US) | 18 |
| | Hebrew (live in Israel) | 16 |
| **Education level** | High-school | 6 |
| | BSc | 18 |
| | MSc | 5 |
| | Phd | 5 |
| **Average age** | | 29.7 |

ter the game started recording while some subjects finished stating their claim before the game finished recording. We removed silence segments with length longer than 300 milliseconds.

We propose three different deception detection models:

1. A Convolutional Neural Network (CNN) model. This model is trained on the spectrograms obtained from the samples. These spectrograms are obtained by transforming the audio samples to images using the Fast Fourier Transform (FFT). The CNN model uses four convolution layers with 3x3 kernels followed by a max-pooling layer with a pool size of 2x2 and finally three fully connected layers with dropout layers between them.

2. A Convolutional Recurrent Neural Network (CRNN) model. This model first uses a single convolution layer on the spectrograms with 3x3 kernel, which is followed by two LSTM cells and a fully connected layer with a softmax activation function.

3. Our multi layer perceptron with Five Sound Feature Model (FSFM). Our neural network consists of 4 fully connected layers with ReLU activation and dropout after each layer. The final layer's uses a softmax activation. It uses a categorical cross entropy loss function and the ADAM optimizer. The five sound features are:

(a) Mel-frequency cepstral coefficients (MFCC), using a 20 ms audio frame unit [23].

(b) Mel-scale spectrogram: A spectrogram in which the frequencies are converted to the mel scale. closer to human performance

(c) Spectral contrast: The difference in amplitude between the spectral peaks and valleys. Spectral contrast can be used to highlight regions of the frequency spectrum [24].

(d) Short-time Fourier transform (STFT): A Fourier transform that takes place around a short time and evaluates the Fourier return on the time-dependent segment. It provides the information regarding fluctuations in frequency contents over time [25].

(e) Tonnetz: A tonal space representation introduced by Euler, [26]. Most literature using Tonnetz features are focused on applications for music [27, 28]. We use the method described in [28], for computing the tonal centroid features.

# 5. Results

## 5.1. Model comparison

We tested all 3 models using 5 fold cross validation on our collected data-set. In addition, we compare our model to the sound base model from [18] named $\text{MLP}_c$. We compared 4 parameters: Accuracy, precision, recall and F1 score. The results can be seen in Table 2. The results clearly show that FSFM outperforms all other models. We therefore select the FSFM model as our deception detection method. FSFM is compared to human performance, and is used as a component in an agent that plays the cheat game against human participants (see Section 6).

Table 2: *A comparison between the performance of the three model candidates, $MLP_c$ from [18], and human players within the scope of the game and outside the game scope.*

| | **Accuracy** | **Precision** | **Recall** | **F-score** |
|---|---|---|---|---|
| **CNN** | 0.587 | 0.382 | 0.265 | 0.313 |
| **CRNN** | 0.60 | 0.52 | 0.42 | 0.494 |
| **FSFM** | **0.665** | 0.563 | **0.529** | **0.546** |
| **$\text{MLP}_c$ [18]** | 0.60 | **0.615** | 0.363 | 0.455 |
| **Humans (in game)** | 0.62 | 0.42 | 0.27 | 0.328 |
| **Humans (outside)** | 0.505 | 0.36 | 0.443 | 0.397 |

## 5.2. Comparison to Human Performance

We first compare the FSFM model against human performance in the game scope. That is, a player calling a cheat is considered as claiming that a sentence is false, while a player playing a different move is considered as claiming that a sentence is true. The results can be seen in Table 2.

However, evaluation within the game scope is problematic, as players may not always call a cheat even if they think their opponent is cheating or vice versa due to various reasons related to the game (a good move to make, no card to play, etc.). We therefore executed another experiment in which the audio samples from the game were played to human subjects outside the scope of the game, and each subject was asked to determine whether a statement sounds true or false. 42 subjects were recruited for this experiment (23 females and 19 males), tagging 943 audio samples. Outside of the game scope the subjects reached a lower accuracy (only 0.5058), but a higher recall (0.4438), see Table 2. The subjects outside of the game scope also obtained a higher precision and F1 score. We believe that this is because that within the game scope, the players were less likely to claim a cheat, due to the game consequences of doing so. A chi-square test shows that FSFM significantly outperforms humans both in the game and out of the game scope ($p < 0.05$).

## 5.3. Predicting Human perception

In addition to the effort for deception detection, our collected data allows us to build a model for predicting whether a human will perceive a given statement as deceptive outside the game scope. Using the same FSFM network construction as in Section 4 we ran the network for 1000 epochs over the data-set using 5 fold cross validation. The model reached an accuracy of 0.5414, precision of 0.4922, recall of 0.446 and an F1 score of 0.4679. We believe that these results can be improved and leave it for future work to develop models that better suit the problem of predicting human perception. We note that the other baseline models that appear in Section 4 obtained lower results.

## 5.4. Multilingual cross training and validation

Since we have high quality labeled data in 2 different languages it was interesting to explore the cultural differences as it comes to lie prediction. In this experiment the data was split according to the language of the sample. We composed four models, two

models were trained and tested on statements from the same language (one in English and one in Hebrew). Two other models were trained on one language and tested on the other. As expected, the models performed better when trained and tested on the same language, than when trained on one language but tested on the other. Furthermore, the model trained on statements from the same language outperformed the model that was trained on both languages and tested on both (the original model, see Table 2). In future work we intend to further investigate this result and attempt to run a language classifier before feeding the audio sample to the FSFM model.

# 6. Deception Detection Agent for the Cheat Game

In this paper we demonstrate the possibility of an autonomous agent working in a deceptive voice based environment. We introduce our Cheat game Autonomous Player Deception Detection Agent (CAPDA). CAPDA uses the predefined model introduced in Section 4 to analyze the voice sample from the human player and decides whether to call a cheat based on the model's evaluation. In addition, it plays any cards it has, but if it does not have appropriate cards it randomly decides whether to make a move with improper cards or to take three cards. The full algorithm is presented at Algorithm 1. The agent uses a pre-recorded set of all the possible claims.

**Result:** Action to be played by CAPDA.
**if** *Agent turn* **then**
  **if** *Possible to call a cheat & ModelEvaluatedAsLie()* **then**
    call a Cheat
  **end**
  **if** *Possible to call a cheat & Opponent is out of cards* **then**
    call a Cheat
  **end**
  **if** *Agent has a legal move* **then**
    drop all cards possible
  **end**
  **if** *unused deck not empty & RandomDouble(0,1)* $\geq$ *0.8* **then**
    take 3 cards
  **else**
    lie and randomly drop 1-2 cards.
  **end**
**end**

**Algorithm 1:** The algorithm used by the Cheat game Autonomous Player Deception detection Agent (CAPDA) to determine which action to take.

For a baseline we use a degenerated version of CAPDA, which we term the "simple agent". The simple agent's algorithm is identical to the CAPDA algorithm with the exception of line number 2. Instead of activating FSFM, it generates a random decision and calls for a cheat 30% of the time. This number was chosen since humans tended to call a cheat on roughly 30% of the statements.

### 6.1. Agent Evaluation

We ran both CAPDA and the simple agent for 40 games against new human players recruited using Mechanical Turk. The results can be seen in Table 3. As depicted in the table, CAPDA was much closer to human performance with a winning rate of 42.5% for CAPDA (i.e., the humans won $57.5\%$ of the games against CAPDA), while the simple agent won only 20% of the cases (i.e., the humans won $80\%$ of the games against the simple agent). These differences are statistically significant ($p < 0.05$; using the chi square test).

Table 3: *Performance of the Agents VS a human player.*

| | **Simple agent** | **CAPDA** |
|---|---|---|
| **Games Played** | 40 | 40 |
| **Games Won** | 8 | 17 |
| **Games Lost** | 32 | 23 |
| **Winning Rate** | 20% | 42.5 % |

Recall that the only difference between the simple agent and CAPDA is in their decision whether to call a cheat on the opponent, where CAPDA uses the deception detection model and the simple agent performs a random decision. Moreover, the decisions made by both agents (when to cheat, what cards to drop, etc.) are quite naïve. Despite that, a great difference in performance is observed between the agents, and CAPDA does not fall far behind human performance. This is an important achievement, showing that an ability to effectively detect deception, substantially increases the overall performance in this game.

# 7. Conclusions & Future work

In this paper we take a step towards the development of agents that can interact with humans in a deceptive speech based environment. We develop a game that allows us to collect high quality voice data of false and true statements given by human subjects. We train a neural network and show that our model has a higher accuracy and overall performs than humans. We built an autonomous agent capable of playing against human opponents in a deceptive environment, and show that an agent using our model of deception significantly outperforms an agent not using this model.

Future work will be dedicated to improving our CAPDA. We will collect more data to be released to the community and improve our models. Another layer of learning will be added to the agent; it will learn its current opponent and improve the model as the game proceeds. The agent will also take strategic actions based on the board state. Another topic for future work is the synthesizing of deceptive speech, i.e., speech that may cause the opponent to believe that an agent is providing a false claim and call a cheat, despite the agent being truthful (or vice-versa).

We will also attempt to apply the methods used in this paper to the pirate game (see [20]). The pirate game is a deceptive environment that allows players to interact with each-other by textual input in a controlled environment. Adding speech to the pirate game will allow an agent to utilize FSFM, the deception model developed in this paper, as well as the model we used to predict human perception.

# 8. Acknowledgment

This research was supported in part by the Ministry of Science, Technology & Space, Israel and the Ministry of Science and Technology of Taiwan.

# 9. References


[1] D. Glodstein, S. L. Glodstein, and J. Fornaro, "Fraud trauma syndrome: The victims of the bernard madoff scandal," *Journal of Forensic Studies in Accounting and Business*, vol. 2, pp. 1–9, 2010.

[2] A. Lavorgna and A. Di Ronco, "Fraud victims or unwary accomplices? an exploratory study of online communities supporting quack medicine," *The many faces of crime for profit and ways of tackling it*, pp. 297–324, 2017.

[3] G. D. R. Team, "A world of lies," *Journal of cross-cultural psychology*, vol. 37, no. 1, pp. 60–74, 2006.

[4] P. Ekman and W. V. Friesen, *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, 2003.

[5] A. Dechêne, C. Stahl, J. Hansen, and M. Wänke, "The truth about the truth: A meta-analytic review of the truth effect," *Personality and Social Psychology Review*, vol. 14, no. 2, pp. 238–257, 2010.

[6] P. V. Trovillo, "History of lie detection," *Am. Inst. Crim. L. & Criminology*, vol. 29, p. 848, 1938.

[7] J. Widacki, "John augustus larson (1892–1965)," *European Polygraph*, vol. 14, no. 1, pp. 9–20, 2020.

[8] V. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, and F. Menczer, "The darpa twitter bot challenge," *Computer*, vol. 49, no. 6, pp. 38–46, 2016.

[9] U. Arsenijevic and M. Jovic, "Artificial intelligence marketing: Chatbots," in *2019 International Conference on Artificial Intelligence: Applications and Innovations (IC-AIAI)*. IEEE, 2019, pp. 19–193.

[10] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception." *Psychological bulletin*, vol. 129, no. 1, p. 74, 2003.

[11] R. Mihalcea and S. Pulman, "Linguistic ethnography: Identifying dominant word classes in text," in *Computational Linguistics and Intelligent Text Processing*, A. Gelbukh, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 594–602.

[12] B. de Ruiter and G. Kachergis, "The mafiascum dataset: A large text corpus for deception detection," *arXiv preprint arXiv:1811.07851*, 2018.

[13] V. Pérez-Rosas and R. Mihalcea, "Experiments in open domain deception detection," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1120–1125.

[14] J. Synnott, D. Dietzel, and M. Ioannou, "A review of the polygraph: history, methodology and current status," *Crime Psychology Review*, vol. 1, no. 1, pp. 59–83, 2015.

[15] H. Nasri, W. Ouarda, and A. Alimi, "Relidss: Novel lie detection system from speech signal," *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, pp. 1–8, 2016.

[16] F. Enos, "Detecting deception in speech," *Ph.D. dissertation, Graduate School Arts Sci., Columbia Univ., New York City, NY, USA*, 2009.

[17] J. B. Hirschberg, S. Benus, J. M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis *et al.*, "Distinguishing deceptive from non-deceptive speech," 2005.

[18] G. Krishnamurthy, N. Majumder, S. Poria, and E. Cambria, "A deep learning approach for multimodal deception detection," *arXiv preprint arXiv:1803.00344*, 2018.

[19] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 59–66.

[20] A. Azaria, A. Richardson, and S. Kraus, "An agent for deception detection in discussion based environments," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2015, pp. 218–227.

[21] W. Hancock, M. W. Floyd, M. Molineaux, and D. W. Aha, "Towards deception detection in a language-driven game." in *FLAIRS Conference*, 2017, pp. 388–393.

[22] J. Bachenko, E. Fitzpatrick, and M. Schonwetter, "Verification and implementation of language-based deception indicators in civil and criminal narratives," in *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2008, pp. 41–48.

[23] M. Xu, L.-Y. Duan, J. Cai, L.-T. Chia, C. Xu, and Q. Tian, "Hmm-based audio keyword generation," in *Pacific-Rim Conference on Multimedia*. Springer, 2004, pp. 566–574.

[24] W. Nogueira, T. Rode, and A. Büchner, "Spectral contrast enhancement improves speech intelligibility in noise for cochlear implants," *Journal of the Acoustical Society of America*, vol. 139, no. 2, pp. 728–739, 2016.

[25] K. Gröchenig, *Foundations of Time-Frequency Analysis*. Boston, MA: Birkhäuser Boston, 2013.

[26] L. Euler, *Tentamen novae theoriae musicae*. Hartknoch, 1739.

[27] E. J. Humphrey, T. Cho, and J. P. Bello, "Learning a robust tonnetz-space transform for automatic chord recognition," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 453–456.

[28] C. Harte, M. Sandler, and M. Gasser, "Detecting harmonic change in musical audio." ACM, 2006, pp. 21–26.