ARIEL UNIVERSITY

MASTER THESIS

Deception Detection by an Autonomous Agent Based on Speech

Author:
Evgeny HERSHKOVITCH NEITERMAN

Supervisor:
Dr. Amos AZARIA

Department of Computer Science

August 6, 2020

Declaration of Authorship

I, Evgeny HERSHKOVITCH NEITERMAN, hereby declare that this thesis entitled, "Deception Detection by an Autonomous Agent Based on Speech" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

Dedicated to my great-grandmother Dora Koichu. Who granted me the love for learning.

Acknowledgements

I would like to thank Ariel University for the fellowship which supported me while conducting my research.

I would like to thank Dr. Amos Azaria for the guidance conducting this research.

I would like to thank my family. My beloved and supportive wife Ronli, without her non of this was possible and my son Reshef.

This work was supported in part by the Ministry of Science and Technology of Israel.

Contents

Ariel University

Abstract

Faculty of Natural Sciences
Department of Computer Science

Master

Deception Detection by an Autonomous Agent Based on Speech

by Evgeny HERSHKOVITCH NEITERMAN

Deception has been around since the beginning of language. Sometimes used for good, but other times used to manipulate people into getting one's way of action. In our research we developed a multilingual autonomous agent that interacts well in a deceptive environment lead by humans. In the first part of this thesis we introduce a controlled environment for large scale, high quality and labeled data collection. The environment is based on an online card game where deception is a requirement for winning. Next, we performed a large scale data collection based on our environment. We have collected over a thousand labeled voice samples from human test subjects, both in English and Hebrew. The data we have collected have been released to the community for further research. In the second part of this thesis, we developed methods for detecting deception based on speech input. In addition, we present models for detecting which statements are perceived as a lie by human participants. Furthermore, we explore the factors of languages and cultures in deception detection. Finally, we have developed autonomous agents that interact with humans in this environment; we show that this agent reaches near-human performance, by using the developed model. Our methodology includes the use of machine learning, natural language processing, image classification and voice analyses methods, for modeling human behavior and intent. These methods are then used by the developed autonomous agent.

List of Figures

List of Tables

List of Abbreviations

CADDA	**C**heat game **A**utonomous **P**layer **D**eception **D**etection **A**gent
VAD	**V**oice **A**ctivity **D**etection
MSE	**M**ean **S**quared **E**rror
NLP	**N**atural **L**anguage **P**rocessing
RNN	**R**ecurrent **N**eural **N**etwork
MLP	**M**ulti **L**ayer **P**erceptron
CRNN	**C**onvolutional **R**ecurrent **N**eural **N**etwork

Chapter 1

Introduction

1.1 Problem of Interest

In this work we develop methods for lie detection based on speech cues using novel methods. From the scientific aspect, we propose a novel method that uses a ranking on the truthfulness of each statement in order to achieve better results. In addition, we develop a model that will detect whether a statement will be perceived as a lie or not by humans. Finally, based on these models we developed agents that interact with humans in deceptive environments. Therefore, the scientific contributions cover comprehensively the upstream technical innovation, the deception behavior discussions, i.e., delivering and perceiving of lies, and the downstream application to the autonomous agents.

1.2 Motivation

Throughout history people have tried to develop a method for lie detection.Throughout Modern history and even further back, many cruel methods were used to detect liars (see [22] for several examples of such methods). In 1921 John Augustus Larson invented the polygraph, a device intended to detect a lie by recording several body measurments, such as breathing rate, pulse, blood pressure, and perspiration. It is assumed that all these measures accelerate while telling a lie. However, the accuracy of the polygraph and similar devices is highly debatable [18, 8, 10], furthermore, these devices require the suspect to be attached to different appliances and cannot be performed retrospectively, or when the suspect is not present. We therefore suggest a method for gathering data that will assist in building human deception models, and finally the development of autonomous agents.

It is hard to overestimate the damage and harm caused by deception and fraud. The Bible states (Leviticus 19,11) "Do not lie, do not deceive one another," and indeed throughout the history, deception has caused the loss of lives and property. However, not all lies may be harmful, and at times, it may be considered wise to tell a lie in order not to avoid hurting one's feeling or similar situations. We believe any intelligent agent must be able to interact in an environment in which humans do not always tell the truth.

We present the development of a multilingual deception detection model based on speech. We developed a game for collecting a large scale and high-quality labeled data-set in a controlled environments in English and Hebrew. We developed a model that can detect deception based only on a vocal statement from the experiment participants, and showed that it performs as well as humans in detecting deceptive speech. We intend to embed the learned model in an autonomous agent that will have a goal to assist human users in avoiding being deceived. In addition,

we developed a model that when, given a vocal statement determines whether it is perceived as deceptive by humans.

As for the social and economic aspects, as far as we know the deception detection technology is in great demand for security companies. They intend to use it in their interviews, and thus we expect that the outcome of this cooperative project to be attached to the market and bring new revenue. Most importantly, we are aware that **there has been some business of the lie detection prototype systems between Israel and Taiwan**. Hence, we believe this academic cooperative project shall lead the way to provide an advanced technical support to the industry and encourage substantial interactions between two countries, which is definitely the final goal of the project call.

1.3 Our Contribution

Developing agents that interact with humans is not simple, especially in deceptive environments. Research into humans' behavior has found that people often deviate from what is thought to be rational behavior, since they are affected by a variety of (sometimes conflicting) factors: a lack of knowledge of one's own preferences, the effects of the task complexity, framing effects, the interplay between emotion and cognition, the problem of self-control, the value of anticipation, future discounting, anchoring and many other effects [23, 15, 1, 6].

Several works have demonstrated that a machine-learning approach, which builds upon psychological factors and human decision-making theory, is essential for developing a good model of true human behavior. The human behavior model is in turn required for successfully implementing agents that interact with humans [9, 14, 21, 19]. In several previous works done in the deep learning lab at Ariel University, the researchers had modeled human behavior by recruiting human subjects via crowd-sourcing platforms and allowing them to interact with a game [4, 3, 2, 16]. Games provide a controlled environment and are a good source for obtaining high quality labeled datasets. We will follow this approach when developing autonomous agents for deceptive environments.

Our expected contributions from this proposal are:

1. The gathering of high-quality datasets for deception detection: (i) A speech data-set with accurate labels. Including the way the statements were perceived by other humans.

2. The development of a lie detector, which uses verbal cues.

3. The development of a ranking based method, which uses ranked (continuous) labels on sentences in order to achieve higher accuracy.

4. The development of a component that detects whether a statement will be perceived as a lie or not by humans.

5. The development of autonomous agents that interact with humans. This agent will use the model developed in the previous phases. We expect to show that the autonomous agents will outperform humans.

Chapter 2

Data Collection

2.1 Collection environment

In this thesis 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 get rid of all of their cards. After dealing eight cards 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. 2.2.

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]. In our local simulation process we found that a two-player game might take too long. Therefore, to prevent players from losing interest and leaving mid-game we limited each game to 12 minutes. This resulted in a data set that is not only labeled by whether each statement is deceptive or not, but also an indication of whether the other opponent thought the statement was deceptive. See Figure 2.1 for a screenshot of the game.

2.2 Collection process

We recruited two types of subjects. US subjects were recruited using Amazon's Mechanical Turk service. These subjects played the game in English. The subjects played the game from their respective computers, from their respective locations, using their respective audio equipment. This resulted in a larger number of low quality recorded samples, which we eventually cleaned. In addition, we gathered graduate students from the Computer Science Department of Ariel University in Israel. The Israeli subjects played the game in Hebrew. Most of the Hebrew-played games, were conducted in our lab at Ariel University using our audio equipment. Fewer samples had to be dropped due to the use of higher quality recording equipment. Each player played three games, where every game ended either when one

[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.

FIGURE 2.1: The graphical interface developed for the "cheat game".

of the players dropped all of their cards or when the 12 minute [2] countdown clock reached zero.

The data set was collected from 34 test subjects. 18 played in the English language and 16 played in Hebrew. Full statistics can be found in table 2.1.

TABLE 2.1: Test subjects details

Parameter	Value	Comments
Number of subjects	34	
Male	18	
Female	16	
Students	10	
Played in English	18	Born in the US
Played in Hebrew	16	Born in Israel
Average age	29.7	
Education level	High-school : 6	
	BSc : 18	
	MSc : 5	
	Phd : 5	

[2]Due to the fast response of the autonomous agents, the games involving an agent were played for up to 8 minutes.

FIGURE 2.2: Example of possible claims, after the last claim was 7(or multiple 7's).

2.3 Data set

We removed noisy data caused due to technical difficulties such as a broken microphone or recordings reported as incomplete by the player's opponent. 177 such recordings were removed. In total, the data collection phase provided 950 labeled samples. 598 true statements and 352 false statements. See Table 2.2 for a summary of the data set. We note that the ratio of the false statements (37%) is more balanced than the ratio in other common deception data sets [5, 20].

TABLE 2.2: Data set Distribution

Group	Samples	Comments
True statements	598	cards matched the statement approved by the opponent
False statements	352	cards did not match statement approved by the opponent
English true statements	327	
English false statements	200	
Hebrew true statements	271	
Hebrew false statements	152	
English statements	527	
Hebrew statements	423	
Total	950	

Chapter 3

Preliminary Model and Results

This chapter describes the work done for a paper called "Multilingual Deception Detection by Autonomous Agents" [13] presented at 2020 WWW conference workshop. It was published midway through this thesis research. The data set and the results described in this chapter are eclipsed by the results in the chapters to follow.

3.1 Data set

The data set at the time contained 637 samples and 26 test subjects. Full statistics of the data set as it was at the time of the papers submission can be found at tables 3.1 3.2.

TABLE 3.1: Data set Distribution as for the WWW2020 conference

Group	Samples	Comments
True statements	395	cards matched the statement approved by the opponent
False statements	242	cards did not match statement approved by the opponent
English true statements	124	
English false statements	90	
Hebrew true statements	271	
Hebrew false statements	152	
English statements	214	
Hebrew statements	423	
Total	637	

3.2 Model

An illustration of the classifier model we developed appears in Fig 3.1. It consists of a voice activity detector, followed by a spectrogram transformation algorithm and the tested algorithm. We describe each component of the model.

- Voice activity detector: There are a number of pauses and silences in the speech samples. These were often due to the fact that some of the speakers did not start speaking at the exact moment the recording started or finished speaking before the recording stopped. We first used voice activity detection (VAD) to

TABLE 3.2: Test subjects details as for the WWW2020 conference

Parameter	Value	Comments
Number of subjects	26	
Male	15	
Female	11	
Students	11	
Played in English	11	Born in the US
Played in Hebrew	15	Born in Israel
Education level	High-school : 5	
	BSc : 15	
	MSc : 4	
	Phd : 2	

remove the silence periods. The VAD threshold was adjusted to match the noise level of the speech samples and we only removed the detected silence segments with lengths longer than 300 milliseconds.

- Spectrogram Transformation Algorithm: this is an open source code provided by matplotlib. It transforms the WAV files spectrogram image based on time and the frequencies in the WAV file (see figure 3.2 for an example). All the spectrograms were 64X64 gray scale images.
- Deep Learning Algorithm: A recurrent neural network with 64 hidden neurons constructed of a LSTM cells. Each mini batch consisted of 64 samples.
- Optimizer and loss function: We used ADAM optimizer with 0.0001 learning rate step. The loss function was MSE.

3.3 Simulation Results

3.3.1 Deception Detection

The RNN described in 3.2 was tested in the following way. We ran a 5-fold test meaning we shuffled the data and split it into 5 groups. We ran the model 5 times, each time a different group works as the test set and the other as the training set. 576 samples were used as the training set and the other 144 as the test set. The network ran for 300 epochs over the train set and was tested against the test set. The results can be seen in Table 3.3.

	True prediction	**False prediction**
True statement	272	100
False statement	155	110

TABLE 3.3: Lie prediction in mixed language train and test

$$Accuracy = 60\%$$

$$Precision = 52\%$$

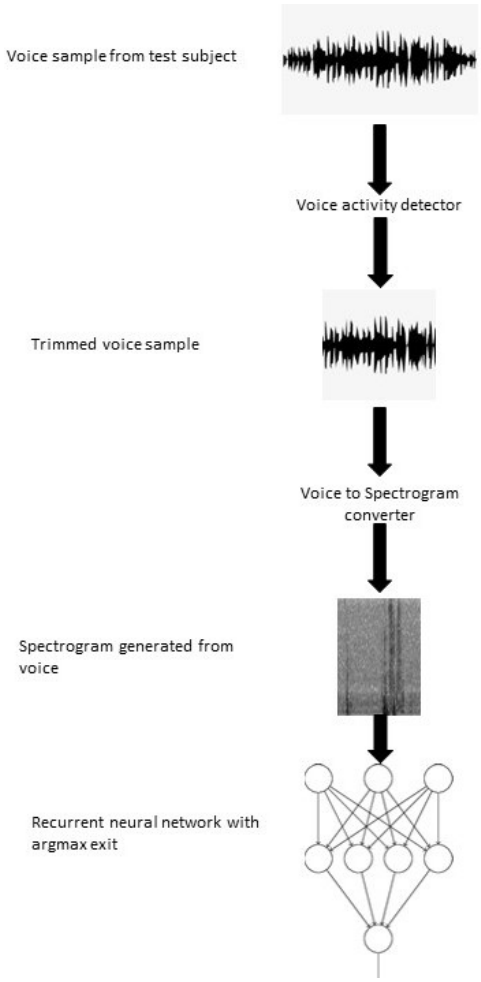

FIGURE 3.1: Illustration of the classification process.

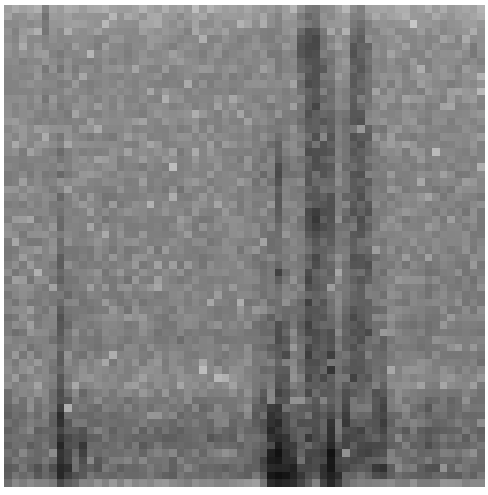

FIGURE 3.2: A visual 64x64 representation of a voice recording from a test subject.

$$Recall = 42\%$$

$$F1score = 46.4\%$$

3.3.2 Comparison to Human Performance

We compared our model against human performances. Table 3.4 presents the results of the human subjects. While the accuracy and precision are relatively similar, the automated model outperforms humans in the Recall rate by 40% and 20% in the F1 score. That is, our model caught significantly more lies than humans ($p < 0.05$; using the chi-square test).

	Human believe	**Human disbelieve**
True statement	334	61
False statement	169	73

TABLE 3.4: Human performance on detecting false statements.

$$Accuracy = 64\%$$

$$Precision = 54\%$$

$$Recall = 30\%$$

$$F1score = 38.5\%$$

3.3.3 Predicting Human perception

In addition to the effort for deception detection, we tried to predict whether a human will perceive a given statement as deceptive. In other words, we tried to predict the behavior of the opponent. We note that there are many other factors involved in the Cheat Game that may cause a player to call their opponent a cheater or play the next turn even though they suspect that their opponent provided a false statement (e.g. no card to play, time running out, good move to make, etc.). Nevertheless, our model managed to get fair results. Using the same RNN network construction as in Section 3.3.1 we ran the network for 1000 epochs over the data set using a 5-fold test. Results can be seen in Table 3.5.

	Predicted human disbelieve	**Predicted human believe**
Human disbelieved claim	30	87
Human believed claim	65	455

TABLE 3.5: Predicting how a human subject will perceive the statement

We note that the data is not balanced, since most of the human subjects believed their opponent's claims. Therefore, the precision, recall and F1 score are much lower when attempting to predict the human subject's perception.

$$Accuracy = 76\%$$

$$Precision = 31\%$$

$$Recall = 26\%$$

$$F1score = 28.2\%$$

3.3.4 Multilingual cross training and validation

Since we have high quality labeled data in two different languages it was interesting to explore the cultural differences when it comes to lie prediction. In this experiment the data was split according to the language of the sample. The model ran once with the English data as the training set and the Hebrew as test set and again when the training and test sets switched places. Results can be found in tables 3.6 3.7.

	True prediction	**Lie prediction**
True statement	193	78
False statement	107	44

TABLE 3.6: Performance of the model when trained on the English data sample and tested on Hebrew.

$$Accuracy = 56\%$$

$$Precision = 36\%$$

$$Recall = 29\%$$

$$F1score = 32.1\%$$

	True prediction	**Lie prediction**
True statement	93	31
False statement	58	31

TABLE 3.7: Performance of the model trained on Hebrew data sample and tested on English.

$$Accuracy = 58\%$$

$$Precision = 50\%$$

$$Recall = 34\%$$

$$F1score = 40\%$$

3.4 Discussion

Analyzing the precision factor of our model compared to human performance shows that humans are more trusting. Human test subjects choose to believe their opponent 79% of the time as opposed to our model that classifies only 67% of the instances

as true. This results in better performance in terms of Precision. The tendency of humans to believe statements told by other people (especially when repeated more than once), is a known effect, that was first identified in 1977 [7].

Chapter 4

Prepossessing and Models

4.1 Prepossessing

4.1.1 Manual cleaning

We built a script to scan the game logs. The Script tagged the samples reported by the players as illegal (according to the Cheat game rules) or missing. Then we manually scanned the remaining samples to detect unclear recordings, which cannot be understood. This resulted in dropping 177 samples.

4.1.2 Voice activity detection

The remaining samples were sent to a Voice Activity Detector (VAD) to trim the recording of silence and background noise. There are a number of pauses and silences in the speech samples. These were often due to the fact that some of the speakers did not start speaking at the exact moment as the recording started or finished speaking before the recording stopped. The VAD threshold was adjusted to match the noise level of the speech samples and we only removed the detected silence segments with lengths longer than 300 milliseconds.

4.1.3 Spectrogram

The spectrogram transformation algorithm is a popular technique used in audio classification, which instead of working directly with the audio sample, the audio is transformed to an image using a Fourier transformation. To save processing time we used Fast Fourier transformation, which generates an acceptable image with a complexity of $O(nlog(n))$ instead of a complexity of $O(n^2)$ for the standard Fourier transformation. An example of a spectrogram is shown in Figure 3.2. We used an open source library to create the spectrograms.

4.1.4 Scaling

Standardization of a dataset is a common requirement for many machine learning estimators: they might malfunction if the individual features do not look similar to standard normally distributed data (e.g. Gaussian with 0 mean and unit variance).

For instance many elements used in the objective function of a learning algorithm (such as the RBF kernel of Support Vector Machines or the L1 and L2 regularizers of linear models) assume that all features are centered around 0 and have variance in the same order. If a feature has a variance whose orders of magnitude are larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected.

We used standard scaling, calculated in the following way:

$$x' = \frac{x - \overline{x}}{\sigma}$$

Where x is the original feature vector, $\overline{x} = \text{average}(x)$ is the mean of that feature vector, and σ is its standard deviation.

4.1.5 Sound feature extraction

For the MLP architecture we inserted several features extracted from different algorithms applied on the voice sample. A visual representation of the extracted features appears in Figure 4.1.

Mel-frequency cepstral coefficients (MFCC)

The mel-frequency cepstrum is highly effective in audio recognition and in modeling the subjective pitch and frequency content of audio signals. Mel scale is computed by:

$$Mel(f) = 2595log_{10}(1 + \frac{f}{700})$$

where $Mel(f)$ is the logarithmic scale of the normal frequency scale f. Mel scale has a constant mel-frequency interval, and covers the frequency range of 0 Hz - 20050 Hz. The Mel-Frequency Cepstral Coefficients (MFCCs) are computed from the FFT power coefficients, which are filtered by a triangular band pass filter bank. The filter bank consists of 12 triangular filters. The MFCCs are computed by:

$$c_n = \sqrt{\frac{2}{K}} \sum_{k=1}^{k} (logS_k) cos[n(k - 0.5)\pi/k], n = 1, 2, ..., N$$

where $S_k(k = 1, 2, ...k)$ is the output of the filter banks and N is total number of samples in a 20 ms audio unit. [24]

Mel-scaled spectrogram

A mel spectrogram is a spectrogram where the frequencies are converted to the mel scale. That is, the frequencies are converted to mel scale using formula 4.1.5, and then the FFT algorithm is used to generate a spectrogram.

Spectral contrast

The difference in amplitude between spectral peaks and valleys is called spectral contrast. It is used with the goal to highlight certain regions of the frequency spectrum around important spectral features, such as format frequencies. It has been shown that such enhancement of contrastive changes in the speech spectrum can improve speech intelligibility for hearing impaired people [17].

Short-time Fourier transform (STFT)

To find the frequency spectrum of a signal f at time x, one localizes f to a neighborhood of x and takes its Fourier transform. This leads to the short-time Fourier

transform (STFT). The localization procedure is parametrized by window function g. Then

$$V_g f(x, \omega) = \int_{\mathbb{R}} f(t)\overline{g(t-x)}e^{-2\pi i t\omega} dt$$

$V_g f$ depends linearly on f, and many of its properties (energy preservation, inversion formula) are similar to those of the Fourier transform [11].

Tonnetz

A feature extraction technique presented in [12]. We compute the tonal centriod features.

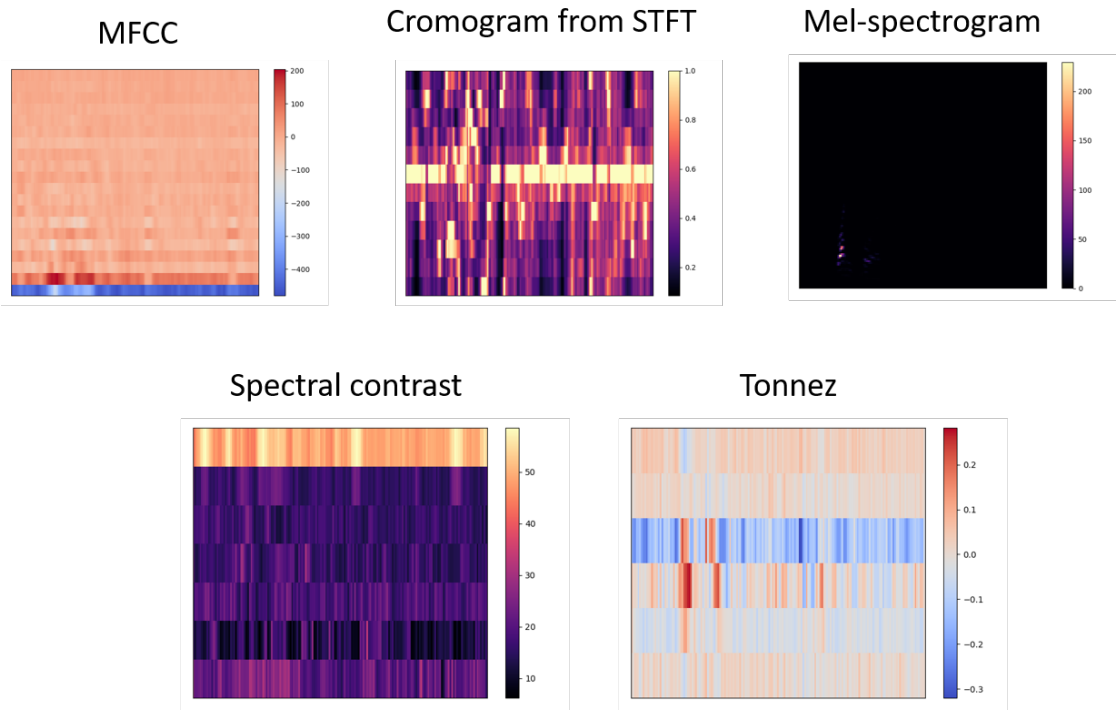

FIGURE 4.1: A visual representation of the sound features extracted from a single sample.

4.2 Models

We propose three different models for the task at hand.

- CNN based on the Fast Fourier transformation (FFT) of the sound sample.
- CRNN based on the same FFT, but with recurrent layers of LSTM cells as part of the hidden layers.
- MLP network with sound feature extraction as a prepossessing step.

All models used Adam optimizer with $learning\_rate = 0.001$

1. CNN:
   Our CNN is constructed with 4 convolution layers with 3x3 kernels. After every convolution layer there is a max pooling layer with a pool size of 2x2

and a dropout layer. After the convolutions there are 3 fully connected layers with dropout layers between them.

2. CRNN:
   Our CRNN is constructed with a single convolution layer with 3x3 kernel followed by 2 recurrent layers from simple LSTM cells. Finally there is a fully connected layer with a softmax activation function.

3. MLP with sound feature extraction:
   Our MLP consists of 4 dense layers with ReLU activation and dropout after each layer. The final layer uses a softmax activation.

4.2.1 Model Comparison

We tested all 3 models using 5-fold cross validation on our collected data set. We have compared 4 parameters: Accuracy, precision, recall and F1 score. The results can be seen in table 4.1. The results clearly show that the MLP model with voice feature extraction (as a prepossessing step) outperforms the other models. We therefore select the MLP model as the model to be used by our agent. Table 4.1 provides the confusion matrix for the MLP model.

TABLE 4.1: Comparison of the three models

	CNN	**CRNN**	**MLP**
Accuracy	58.7%	60%	**66.5%**
Precision	38.2%	52%	**56.3%**
Recall	26.5%	42%	**52.9%**
F1 score	31.3%	49.4%	**54.6%**

4.3 Results

Considering that the MLP appears to be the best performing model as shown in 4.2.1, from this point on we will be using it for the results ahead and as the model of choice for the Smart agent presented at 5.1. We used 5-fold cross validation to evaluate the performance of our model. The network ran for 300 epochs over the training set and was tested against the test set. The results appear in Table 4.2.

TABLE 4.2: Lie prediction in mixed language train and test MLP with voice feature extraction

	True prediction	**False prediction**
True statement	441	148
False statement	170	191

$$Accuracy = 66.5\%$$

$$Precision = 56.3\%$$

$$Recall = 52.9\%$$

$$F1\ score = 54.6\%$$

4.3.1 Comparison to Human Performance

We compare the MLP model against human performances. Table 4.3 presents the results of the human subjects. A chi-square test shows that our model significantly outperforms humans ($p < 0.05$; using the chi-square test). Moreover, all the commonly tested categories show that our model is better than a human in the task of deception detection through the use of the voice samples.

TABLE 4.3: Human performance on detecting false statements.

	Human believe	**Human disbelieve**
True statement	501	119
False statement	241	89

$$Accuracy = 62\%$$

$$Precision = 42\%$$

$$Recall = 27\%$$

$$F1\ score = 32.8\%$$

4.3.2 Predicting Human perception

Beyond the effort for deception detection, we tried to predict whether a human will perceive a given statement as deceptive. Strictly speaking, we tried to predict the behavior of the opponent. We note that there are many possible reasons that a player may choose to call another player a cheater, or continue to play even though it is suspected that their opponent is cheating (e.g. no card to play, time running out, good move to make, etc.). Nevertheless, our model was capable of getting equitable results. Using the same MLP network construction as in Section 3 we ran the network for 1000 epochs over the data set using a 5-fold test. Results can be seen in Table 4.4.

TABLE 4.4: Predicting how a human subject will perceive the statement

	Predicted human believe	**Predicted human disbelieve**
Human believed claim	617	163
Human disbelieved claim	134	36

$$Accuracy = 68.7\%$$

$$Precision = 18\%$$

$$Recall = 21\%$$

$$F1\ score = 19.3\%$$

We have observed that the data is not uniform, as many of the human subjects believed their opponents' claims. Therefore, the precision and recall are much lower when attempting to predict the human subject's perception.

4.3.3 Multilingual cross training and validation

Since we have high quality, labeled data in 2 different languages it was remarkable to discover the cultural differences as when it comes to lie prediction. In this experiment the data was split according to the language of the sample. The model ran once with the English data as the training set and the Hebrew data as the test set and then this process was reversed. Results can be found in tables 4.5 4.6.

TABLE 4.5: Performance of the model when trained on the English data sample and tested on Hebrew.

	True prediction	**Lie prediction**
True statement	118	153
False statement	69	83

$$Accuracy = 47.5\%$$

$$Precision = 54.6\%$$

$$Recall = 35.1\%$$

$$F1\ score = 42.7\%$$

TABLE 4.6: Performance of the model trained on Hebrew data sample and tested on English.

	True prediction	**Lie prediction**
True statement	243	84
False statement	146	54

$$Accuracy = 56\%$$

$$Precision = 27\%$$

$$Recall = 39\%$$

$$F1\ score = 32\%$$

Chapter 5

Autonomous agents

5.1 Deception Detection Agent for the Cheat Game

In this thesis 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 (CAPDDA). CAPDDA uses the predefined model introduced in 3 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 (the autonomous agent) 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 prerecorded set of all the possible claims.

Algorithm 1: CAPDDA Algorithm

```
   Result: decision which type of move to play
1  if Agent turn then
2      if Possible to call a cheat & ModelEvaluatedAsLie() then
3          call a Cheat
4      end
5      if Possible to call a cheat & Opponent is out of cards then
6          call a Cheat
7      end
8      if Agent has a legal move then
9          drop all cards possible
10     end
11     if unused deck not empty & RandomDouble(0,1) ≥ 0.8 then
12         take 3 cards
13     else
14         lie and randomly drop 1-2 cards.
15     end
16 end
```

As a baseline we developed the simple agent, a degenerated version of CAPDDA. The algorithm, is identical to the CAPDDA algorithm with the exception of line number 2. Instead of activating our deep learning model, it generates a random decision and calls for a cheat 30% of the time.

5.1.1 Agent Comparison

We ran both CAPDDA and the simple agent for 40 games against human players. The results can be seen in 5.1. As depicted in the table, CAPDDA was much closer

to human performance with a winning rate of 42.5%, while the simple agent won only 20% of the cases. These differences are statistically significant ($p < 0.05$; using the chi square test).

TABLE 5.1: Performance of the Agents VS a human player.

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

One may recall that the only difference between the simple agent and CAPDDA is in their decision whether to call a cheat on the opponent and where CAPDDA 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 all of that, a great difference in performance is observed between the agents, and CAPDDA 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. This might indicate that embedding our algorithm in deceptive environments, such as business or diplomatic meetings, could give the user an edge over the other side.

Chapter 6

Conclusions

6.1 Discussion

Analyzing the precision factor of our model compared to human performance shows that humans are more trusting. Human test subjects choose to believe their opponent 79% of the time as opposed to our model that identifies only 64.3% of the instances as true. This results in better performance. The tendency of humans to believe statements told by other people (especially when repeated more than once), is a known effect, that was first identified in 1977 [7].

We also note the difference between men and women. Male players provided 176 false claims out of 561 claim (31% of the male claims being a lie). Female players provided 175 false claims out of 389 claims in total (a 45% lie ratio). The differences are statistically significant ($p < 0.05$; using the chi-square test). These differences may be attributed to several aspects including the willingness to take risks. However, no statistically significant differences were found between cultures, with 36% of the statements provided by Israeli subjects being false, compared to 38% of the subjects from the US.

One criticism against our developed model may be that powerful organizations may use our model for lie-detection against individuals, including their employees. However, we note that powerful organizations have access to more invasive methods such as the use of a polygraph, and therefore our model will be of little interest to them. Therefore, our model will be more valuable for individuals when operating in a deceptive environment and might assist them from being deceived and scammed.

6.2 Conclusions & Future work

In this thesis we take a step towards the development of agents that can assist human users, with an attempt to avoid fraud and being misled. 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, outperforms humans. We built an autonomous agent capable of playing against human opponents in a deceptive environment, and showed that an agent using our model of deception significantly outperforms an agent not using this model.

Future work will be dedicated to improving our CAPDDA. 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 thesis to the pirate game (see [2]). 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 the use of the deception model developed in this thesis as well as the model used to detect human perception to an agent interacting in the pirate game.

Bibliography

[1] Dan Ariely, George Loewenstein, and Drazen Prelec. ""Coherent arbitrariness": Stable demand curves without stable preferences". In: *The Quarterly Journal of Economics* 118.1 (2003), pp. 73–106.

[2] Amos Azaria, Ariella Richardson, and Sarit 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.

[3] Amos Azaria et al. "Giving Advice to People in Path Selection Problems". In: *AAMAS*. 2012.

[4] Amos Azaria et al. "Strategic Advice Provision in Repeated Human-Agent Interactions". In: *IJCAI*. 2012.

[5] Joan Bachenko, Eileen Fitzpatrick, and Michael 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.

[6] C. F. Camerer. "Behavioral Game Theory. Experiments in Strategic Interaction". In: Princeton University Press, 2003. Chap. 2.

[7] Alice Dechêne et al. "The truth about the truth: A meta-analytic review of the truth effect". In: *Personality and Social Psychology Review* 14.2 (2010), pp. 238–257.

[8] Paul Ekman and Wallace V Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, 2003.

[9] Ya'akov Gal and Avi Pfeffer. "Modeling reciprocal behavior in human bilateral negotiation". In: *Proceedings of the National Conference on Artificial Intelligence*. Vol. 22. 1. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999. 2007, p. 815.

[10] Don Grubin and Lars Madsen. "Accuracy and utility of post-conviction polygraph testing of sex offenders". In: *The British Journal of Psychiatry* 188.5 (2006), pp. 479–483.

[11] Karlheinz Gröchenig. *Foundations of Time-Frequency Analysis*. English. Boston, MA: Birkhäuser Boston, 2013. ISBN: 0817640223;9780817640224;

[12] Christopher Harte, Mark Sandler, and Martin Gasser. "Detecting harmonic change in musical audio". English. In: ACM, 2006, pp. 21–26. ISBN: 9781595935014;1595935010;

[13] Evgeny Hershkovitch Neiterman, Moshe Bitan, and Amos Azaria. "Multilingual Deception Detection by Autonomous Agents". In: *Companion Proceedings of the Web Conference 2020*. 2020, pp. 480–484.

[14] Koen Hindriks and Dmytro Tykhonov. "Opponent modelling in automated multi-issue negotiation using bayesian learning". In: *AAMAS*. 2008, pp. 331–338.

[15] George Loewenstein. "Willpower: A Decision-theorist's Perspective". In: *Law and Philosophy* 19 (1 2000), pp. 51–76. ISSN: 0167-5249.

[16] Thanh Hong Nguyen et al. "Analyzing the effectiveness of adversary modeling in security games". In: *Twenty-Seventh AAAI Conference on Artificial Intelligence*. 2013.

[17] Waldo Nogueira, Thilo Rode, and Andreas Büchner. "Spectral contrast enhancement improves speech intelligibility in noise for cochlear implants". English. In: *Journal of the Acoustical Society of America* 139.2 (2016), pp. 728–739.

[18] Christopher J Patrick and William G Iacono. "Psychopathy, threat, and polygraph test accuracy." In: *Journal of Applied Psychology* 74.2 (1989), p. 347.

[19] Avi Rosenfeld and Sarit Kraus. "Using aspiration adaptation theory to improve learning". In: *AAMAS*. 2011, pp. 423–430.

[20] Bob de Ruiter and George Kachergis. "The Mafiascum Dataset: A Large Text Corpus for Deception Detection". In: *arXiv preprint arXiv:1811.07851* (2018).

[21] Ventatramanan S Subrahmanian. *Heterogeneous agent systems*. MIT press, 2000.

[22] Paul V Trovillo. "History of lie detection". In: *Am. Inst. Crim. L. & Criminology* 29 (1938), p. 848.

[23] Amos Tversky and Danliel Kahneman. "The Framing of Decisions and the Psychology of Choice". In: *Science* 211.4481 (1981), pp. 453–458.

[24] Min Xu et al. "HMM-based audio keyword generation". In: *Pacific-Rim Conference on Multimedia*. Springer. 2004, pp. 566–574.

תקציר–

הונאה הייתה קיימת טרם תחילת השימוש בשפה. לעיתים שימשה את האדם לטוב, אך בפעמים אחרות היוותה כלי לביצוע מניפולציות והשגת מטרות של האינדיבידואל. במחקרנו פותח סוכן אוטונומי רב-לשוני המסוגל לתקשר עם בני אדם, בסביבה בה ההונאה הינה חלק משמעותי מהתקשורת.

בחלק הראשון של התזה יצרנו סביבה מבוקרת לאיסוף מידע איכותי, מתויג, ובהיקף גדול. פיתחנו משחק קלפים ברשת בו שקרים נדרשים בכדי לנצח. לאחר מכן, ביצענו איסוף מידע בהיקף גדול כאשר הסביבה שבנינו משמשת ככלי לאיסוף המידע. אספנו כאלף דגימות קול משחקנים אנושיים בעברית ובאנגלית. המידע שנאסף שותף עם הקהילה המדעית להמשך מחקר.

בחלקה השני של התזה, פיתחנו טכניקות לגילוי שקרים והונאה על בסיס דיבור. בנוסף אנו מציגים מודלים לניבוי ההתנהגות האנושית, ויכולת חיזוי לשאלה: "האם האדם יזהה את ההצהרה כשקרית?". מעבר לכך אנו חוקרים את הגורמים בשפה ובתרבות בנוגע לזיהוי שקרים.

בסיום עבודתנו אנו מפתחים סוכן אוטונומי רב-לשוני המתקשר עם בני אדם בסביבה זו. אנו מראים כי סוכן זה מגיע לביצועים כמעט אנושיים, כאשר הוא משתמש במודל שפיתחנו.

המתודולוגיות שלנו כוללות שימוש נרחב בלמידת מכונה, עיבוד שפות טבעיות, קטלוג תמונות וטכניקות שונות של עיבוד קול, לצורך מידול ההתנהגות האנושית וכוונתו.

תודות–

- ברצוני להודות לאוניברסיטת אריאל על המלגה האישית אשר שימשה עבורי גב כלכלי במהלך המחקר .
- בנוסף, ברצוני להודות במיוחד לד"ר עמוס עזריה שליווה אותי לכל אורך הדרך.
- ברצוני להודות למשפחתי התומכת, לאשתי האהובה רונלי ולבני רשף.
- עבודה זו נתמכה בחלקה ע"י משרד המדע והטכנולוגיה של ישראל.

אוניברסיטת אריאל בשומרון

הפקולטה למדעי הטבע

גילוי שקרים ע"י סוכנים אוטונומיים מבוססי דיבור.

חיבור זה מוגש כחלק מהדרישות לקבלת התואר "מוסמך האוניברסיטה" (M.Sc.)

במחלקה למדעי המחשב

על ידי:

יבגני הרשקוביץ נייטרמן

העבודה הוכנה בהדרכתו של ד"ר עמוס עזריה

ט"ז באב תש"פ, 6.8.2020