Biometric-enabled Interview-Assisting Traveller Screening technology for Automated Border Control

Technical Report · February 2018

DOI: 10.13140/RG.2.2.29946.82880

2 authors, including:

Dmitry O. Gorodnichy
Government of Canada
266 PUBLICATIONS **1,290** CITATIONS

SEE PROFILE

Design and Evaluation of Biometric-enabled Interview Assisting Traveller Screening Technology

CSSP-2015-TI-2158 Study Full Report

Dr. Dmitry O. Gorodnichy
Dr. Aaron Elkins

The publication of this article is enabled by the Scientific Integrity Policy released by the Canada Border Agency Services (CBSA) in December 2019 and is done in support of the Government of Canada's National Action Plan on Open Government and Open Science.

This article is the unabridged version of the report that was prepared by the authors for the CSSP-2015-TI-2158 Study. The reference to the abridged published version is provided below:

> Dmitry O. Gorodnichy, "*Design and Evaluation of Biometric-enabled Interview-Assisting Traveller Screening technology for Automated Border Control. CSSP-2015-TI-2158 Study . Final Report*", Technical Report, DRDC-RDDC-2018-C223, Online: http://cradpdf.drdc-rddc.gc.ca/PDFS/unc331/p808530_A1b.pdf.

This article should be referenced as follows:

> Dmitry O. Gorodnichy and Aaron Elkins, *"Design and Evaluation of Biometric-enabled Interview-Assisting Traveller Screening technology for Automated Border Control. CSSP-2015-TI-2158 Study . Full Report*", Online:

Disclaimer

Opinions presented in this article are solely those of the author and do not represent the opinion of Government of Canada. The work presented in this article was done in 2015-2017 and may not reflect the latest information related to the subject.

Abstract

This report presents the deliverables for the “Roadmap for Biometrics at the Border" (CSSP-2015-TI-2158) study conducted by the Canada Border Services Agency (CBSA) in partnership with University of Arizona (UA) and San Diego State University (SDSU) through support from the Defence Research and Development Canada, Canadian Safety and Security Program (CSSP). The main objective of this study was to generate critical knowledge related to the use of biometric-enabled Interview Assisting Traveller Screening (IATS) technology, such as AVATAR kiosks developed by UA and SDSU. The deliverables include: overview of manual behaviour screening limitations, overview of challenges related to designing biometric-enabled behaviour screening (BEBS) systems, development of a novel framework for designing and evaluating BEBS systems, conducting a mock-up experiment with the AVATAR kiosk at the CBSA, and finally, recommendations based on the lessons and insights learnt from the conducted evaluations.

About Authors

Dmitry O. Gorodnichy is a senior research scientist with the Canada Border Services Agency , a graduate from the Artificial Intelligence Laboratory at the University of Alberta, formerly a researcher with the National Research Council of Canada.

Aaron Elkins is an Assistant Professor at SDSU, where he is the founding director of the Artificial Intelligence Laboratory.

Acknowledgements

This work was done under the funding from the Canadian Safety and Security Program (CSSP) managed by the Defence Research and Development Canada, Centre for Security Science (DRDC-CSS), in partnership with University of Arizona (UA) and San Diego State University (SDSU). Other members of the project included Judee Burgoon, Elyse Golob, Jay Nunamaker, and Bradley Walls from UA. Their feedback and contributions to this project are gratefully acknowledged and are presented in separate technical papers listed in References.

The authors also gratefully acknowledge the feedback and assistance from the Center for Identification Technology Research (CITeR) and in particular from the CITeR Director, Stephanie Schuckers, who introduced the CBSA scientist to the UA team, which led to the development of this project.

The help of former frontline officers Marie-Eve Boulanger and Moe Mohtadi, in preparing the interview questions for AVATAR, as well as valuable feedback from Kevin Kearney, and help of many other of CBSA employees in organizing the AVATAR workshop and demonstration at the CBSA, is also gratefully acknowledged.

Dedication

This report is dedicated to Diane Keller, the first Director General of the CBSA's Science and Engineering Directorate (2006-2017), whose dedication to science and scientists made it possible for the Agency to become an internationally recognized player in applying Biometric and Video Analytics research to border applications.

Diane Keller with AVATAR during the AVATAR test at the CBSA in March 2016.

Contents

List of Figures

List of Tables

Executive summary

This report presents the outcomes of the two-year research study conducted in partnership with University of Arizona (UA) and San Diego State University (SDSU). The main objective of this study was to generate critical knowledge related to the use of *biometric-enabled Interview Assisting Traveller Screening (IATS*) technology, such as AVATAR kiosks developed by UA and SDSU. Secondary objectives included conducting the analysis of related normative references and the development of practical guidelines for deploying this technology in Canada. These objectives were fully met, with the following outcomes achieved.

1. Recommendation to ISO and novel framework for evaluating biometric-enabled screening systems

According to UA scientists, who have over 40 years of research experience in behavior screening, poor performance in detecting lies affects both novices and professionals, with accuracy of detection ranging from 45% (novices) to 65% (professionals). Automated biometric-enabled behaviour screening *(*BEBS) is seen as a way to address challenges of manual screening. However, there have been no standards developed to date to evaluate and report the performance of such systems. International Standard Organization (ISO), who is responsible for developing standards, has limited the scope of its biometric standards to only those that deal with traditional identification-related applications. As a result, any BEBS performance numbers that have been reported to date may not be considered scientifically validated. No guidelines exist to assist organizations in designing or evaluating BEBS systems. This study addressed this gap.

A list of new biometric terms related to BEBS applications has been developed and submitted to International Standard Organization (ISO) , presented in a separate report, and a novel framework for designing and evaluating such systems has been developed. According to this framework, BEBS are designed and evaluated in a way this is done in video analytics and medical applications, rather than how this is done in traditional (ISO-defined) biometrics. This framework is the most important contribution of the study, which now allows GoC to build and test BEBS systems. It was applied for testing the AVATAR technology, making it possible to identify the biometric features and person actions that elicit the best discriminatory power of the system.

2. Testing of the AVATAR kiosk at the CBSA

In March 2016, through the joint effort of UA and the CBSA, the AVATAR kiosk was brought to the CBSA for a prototype testing exercise. Over eighty volunteers participated in this exercise, making it possible to collect what is now the largest data-set for the research in automated behaviour screening for ABC applications. The dataset contains over a million biometric and non-biometric measurements that were collected during automated two-minute interviews conducted by the AVATAR kiosk. The kiosk recorded person's voice, face and pupil dynamics using a stereo microphone, HD camera, and eye tracker installed inside the kiosk. The dataset was used to analyze, the performance of the AVATAR kiosk and to gain new knowledge related to the ways of further improving the performance of AVATAR-like systems.

In their current state, AVATAR kiosks, are shown to be able to detect deception, with accuracy ranging from 55% to 80%. These numbers however are based on a number of study limitations and assumptions that may not hold for a real-life application: the results are obtained on a very small population size (82 volunteers who participated in the experiment and who were not real travellers or real smugglers); they are obtained on the same dataset that was used to tune the algorithm, which creates an "optimistic" bias in the reported accuracies. Finally, CELP (Cultural, Ethical, Legal and Privacy) constraints were not evaluated, as being out of study scope. Of larger importance than the reported accuracy metrics are the developed methodology and the new insights that have been gained from testing the AVATAR kiosks , as they allow one to further improve the performance of BEBS systems.

3. New insights on manual and automated screening

This study obtained new evidence on the behavioural characteristics that are affected by lying (the so called deception signals). Two types of deception signals are identified: those effected by overall stress due to

performing unlawful actions, and those affected specifically by having to lie. This finding may be found useful not only for automated system developer, but also for human interrogators.

The biometric features and interview questions than elicit the best discriminatory paper of the systems have been also identified. At the same time, it has been also shown that even the best performing biometric modalities (such as pupil, voice, and eye fixation) may not contribute to better lie detection if used incorrectly. This is why it is concluded that extreme scientific rigour and integrity should be exercised prior to deployment of such systems in the field, because of the elevated risk of falsely flagging travellers due to technology limitations and the absence of public standards for the evaluation of such technology.

4. Recommendations for next steps

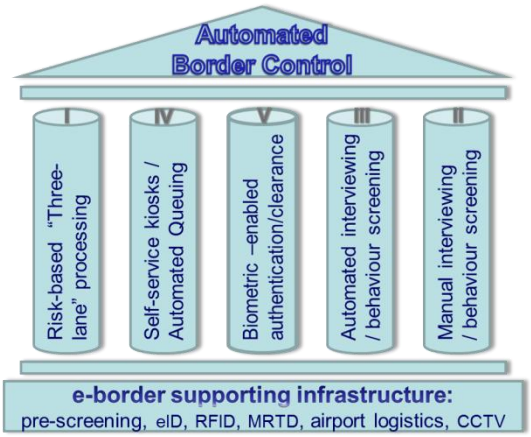
Using the semaphore-like PROVE-IT technology readiness assessment methodology developed in previous projects, the Technology Level Readiness (TRL) of the BEBS systems is accessed as being "yellow", i.e., potentially ready for deployment within next three-five years, provided that adequate in-house research and development capability is available to further tune and test them.

For the agency, the main opportunity is seen in using the no-voice IATS / BEBS as part of the next-generation Primary Inspection Kiosks (PIK). PIK kiosks are already equipped with video and iris cameras. Provided that these cameras are at all times aligned with travellers' faces (so that faces and eyes of all travellers are always clearly seen by the camera), and an additional eye tracking sensor is installed, PIK kiosks should be able to extract biometric signals that can be used for deceit detection.

IATS / BEBS can also be used for manual screening at PIL booths, pre-screening (for trusted travellers applications and staff recruitment), post-screening (i.e., in Secondary examination), or for remote screening (as in self-reporting over a telephone or at remote unmanned ports of entry).

In all of these applications IATS / BEBS may be used in both automated and semi-automated fashion. In the former, the signals detected by sensors are interpreted by the machine, in the latter they are interpreted by humans.

1 Introduction



The precursor "ART in ABC" study (CSSP-2013-CP-1020) [1] has established *five* key border modernization traveller screening components, *Pillars*, that will make automated border control (ABC) of the future:

I. "three-level" risk processing,
II. manual interviewing and behavior screening,
III. automated interview-assisting behaviour screening,
IV. automated queuing / self-service kiosks,
V. biometric-enabled authentication.

As pinpointed in the "ART in ABC" study, even trained officers, as in the US SPOT program have difficulty correctly identifying "suspicious" behaviors. Furthermore, there is enough evidence to suggest that humans, knowingly or unknowingly, will always be vulnerable to "human bias" in making their decisions, due to cultural, religious, gender, and other differences. It appears that, with existing practices and training, manual interviewing and behavior screening has reached the plateau limit in its efficiency.

To address this limitation of human performance in behavior screening, the researchers from University of Arizona, started to develop AVATAR (short for Automated Virtual Agent for Truth Assessments in Real-Time) – the kiosk aimed at automating behaviour screening [3]. The AVATAR and other *interview-assisting traveller screening (IATS)* technologies are seen as a way to scale and further improve the interviewing and behavior screening of incoming travellers.

Following the recommendations from the "ART in ABC" study, the CBSA took the lead to further explore the readiness of IATS technologies for ABC applications. Through support from the Defence Research and Development Canada (DRDC) Canadian Safety and Security Program (CSSP), it established a new project with the objective to generate the critical knowledge related to IATS technologies and to develop practical recommendations related to further testing and deployment of these technologies in Canada.

For the purpose of achieving the objectives, a new partnership with the researchers from the University of Arizona was established. The project ran from September 2015 to December 2017. The outcomes of this project are presented in this report. The report is organized as follows.

Section 2 addresses the limitations of manual screening of travellers. Section 3 provides background behind automated detection of stress and emotions. Section 4 describes the AVATAR technology and the mock-up experiment conducted at the CBSA to demonstrate and test this technology. Section 5 presents the novel methodology that was designed for designing and evaluating AVATAR-like technologies. Results from testing the AVATAR kiosk are presented in Section 6. The concluding section summarizes the limitations and advantages of IATS technologies and presents recommendations related to testing of these technologies in the field.

2 Manual behaviour screening

2.1 Two parts of the challenge

With over a million travellers entering the country daily and over a thousand border officers who have to assess the risks associated with each traveller, the risk of not being able to efficiently validate person's credibility can be high. This section aims at providing the information that may allow the agency to minimize this risk.

Behaviour screening needs to be understood as a *two-sided problem*. The first side of the problem relates to the limitation of human abilities in detecting lies. According to the UA scientists [3,7], who have over 40 years of research experience in deceit detection and behavior screening, poor performance affects both novices and professionals, with accuracy of detection ranging from 45% (novices) to 65% (professionals). Furthermore,

research evidently shows that confidence in judgment is not correlated with accuracy (Correlation < 0.05). The list of human limitations and typical biases and misconceptions related to lie detection is provided in Table 2-1.

The other side of the problem relates to the cultural and mental diversity of humans. As highlighted in the precursor study [1], manual behavior screening may lead to wrong decisions due to human error with respect to individuals who have anxiety or other mental health conditions, the percentage of whom is estimated as 20% of the household population (according to the Canadian Mental Health Association). Furthermore, travellers are commonly already under stress due to travel-related challenges and frequently come from different, possibly unknown, cultural backgrounds, which makes them even more vulnerable to wrong decisions with respect to their behaviour.

Table 2-1 Typical human biases, limitations, and misconceptions related to lie detection.

Typical human biases	Human limitations	Typical misconceptions
• Truth bias - Tendency to assume all tell truth - Common among lay-people • Othello (lie) bias - Tendency to assume all are lying - Common among law enforcement personnel • Cultural / religious bias	• Limited ability to view all signals (e.g. pupil dilation, heart beat) • Limited capacity to analyze multiple cues at a time (normally, less than five) • Attention required for other tasks (watching people, luggage, computer screens, etc.) • Overconfidence, which does not correlate with quality • Prone to misconceptions	– Gaze aversion – Nervous gesturing – Preening

3 Automated behavior screening

3.1 Deception signals and four benefits of computerizing their detection.

Table 3-1 shows the signals that are believed to be related to lying. Computerized recognition of these signals is seen as a way to address challenges of manual screening described above. In particular, the following four benefits are expected from using automated behavior screening for automated border control:

- Improving accuracy (less false hits and less misses),
- Alleviating the "human bias",
- Allowing scalability of screening solutions,
- Making credibility assessment auditable.

The latter will make it possible to know on which grounds, i.e., because of which behaviour signals and actions, the credibility of a traveller is questioned. This may help to improve the service quality.

3.2 Factors contributing the development of the technology

Recognition of person's attributes (such as facial expression, fatigue, motion patterns) from measurable data has been a popular research topic in academia for several decades with over a thousand research papers published on the topic yearly. Over the past decade however this research area has received a particular boost, due to the following three factors:

i) Sensors have become ubiquitous, affordable and diverse. For example, standard smart-phones now have at least five sensors inside: camera, microphone, GPS tracker, proximity tracker and gyroscope tracker,
ii) Many third-party open-source software have been developed to extract various numerical measurements from these sensors,
iii) Many open-source machine learning packages, including deep neural networks, have been developed and available for free and are widely supported

While major advancements have been made in developing BEBS over the past decade, many challenges remained, which are summarized next.

3.3 Stress detection: two tasks of the problem and key questions

To appreciate the complexity of the automated signals detection work, one needs to understand that for a computer, as opposed to a human, the "Detection" is, in fact, a two-task problem. The first task deals with being able to *measure* the signals ("Registration" problem). The second task deals with *interpreting* the measured signals ("Recognition" problem"). While the "Measuring" problem is generally easier for a computer than for a human, the "Recognition" problem may not (see Table 3-1). Therefore, the following research questions are critical for the development of BEBS technology for ABC:

For the "Measuring" problem:

- *Which sensors can be used to capture deception signals ?*
- *Which of these sensors can be used in ABC applications?*

For the "Recognition" problem:

- *How to convert measured raw signals into features?*
- *Which of features are useful for automated recognition of deceit and which are not*?

Table 3-1 Deception signals, sensors that can measure them, level of difficulty in their detection.

Deception signals categorized by biometrics modalities	**for humans to detect**	**sensors that can be used**	**for machine to capture**	**for machine to recognize**
Oculometrics:				
-Pupil size dynamics / change ↑ -Eye movement -Blink patterns	-hard -hard -medium	eye-tracker, video-camera	easy	medium
Kinesic signals:				
Liars are more tense / less expressive (fewer illustrators) -Micro-facial expressions -Body movements (head, hands, legs, torso) • Posture, Stance, proximity • shifts & rigidity • Initial freeze response • Finger fidgeting • Hand to face adaptors • More lip presses	 -medium -medium	video-camera, eye-tracker, gyro sensor (e.g. on tablet in person's hands) weight sensor	easy easy	hard hard
Physiometrics:				
-Body / face temperature ↑ -Brain activity ↑ -Heart rate ↑ -Respiratory patterns ↑	- Hard, impossible	camera, IR-camera, special sensors	medium	hard
Auditory signals - Vocalics:				
Voice Quality↓ -Harmonics-to-noise ratio decreases Pitch, Tempo, Intensity↑ • Fundamental frequency ↑ • change in pitch ↑ • Tension ↑ Response Latency↑ Disfluency↑	-easy-medium -easy-hard (depends on training)	microphones	easy	medium
Auditory signals - Linguistic:				
-What is being said (context, logic, consistency) -How something is being said (sentiment, choice of vocabulary) • Sample message features • Average sentence length • Passive voice ratio, Emotional content, Word diversity	easy-medium	microphones	hard hard	hard easy

↑ / ↓ indicate the increase / decrease of the feature value with increase of stress.

These questions need to be addressed taking into account that, besides technical challenges (such as how to measure the signals under the application and how to build the recognition algorithm), there are also CELP challenges – related to Cultural, Ethical, Legal and Privacy issues, as discussed in our past projects [2,7].

In theory, any number of sensors can be used for BEBS. In practice however one needs to investigate which ones are efficient and which are not. Table 3-1 lists sensors that have been mentioned in literature for the purpose of measuring various biometric signals. Some of these sensors have been used for many years in interrogations using polygraphs (where detected signals are analyzed by a qualified interrogator). Some others have not been used with general public yet and therefore require more intense CELP-related effort to allow the use in the field.

It should be realized that, while in laboratory settings some signals are easier to measure than others, in the operational settings (because of the application constraints), these signals may not be well measured, i.e. with precision required for their recognition.

3.3.1 Academic landscape

What makes stress recognition good (easy) for developers, and unfortunately bad (hard) for its evaluation, is the fact that, unlike classical (ISO-defined) biometrics, it does not actually deal with "recognition" of data (in the machine learning terminology) but rather with the much simplified version of it, known as "classification" or "clustering", where the number of classes (e.g., "stressed" vs. "not-stressed", "happy" vs. "unhappy", etc.) is relatively small and where these classes may not be precisely defined.

Furthermore, it is common for academia and industry to use the third party video and audio processing software, several of which are available for free, to generate features. This software is treated as "black boxes", i.e., without requiring any understanding on how they work. The generated features are then plugged into another "black box" third party software, such as various machine learning (a.k.a. data science) techniques available in R [9], again without the need for a deep understanding on how these techniques work. As a result, it becomes relatively simple for developers, even without knowledge in Artificial Intelligence (AI) , to develop an AI system for stress (or any other type of behavioural indicator) recognition, and it is very difficult for end user to evaluate the quality of such systems.

The best approach to proceed with developing BEBS systems is seen in partnering with academic organizations working in AI who do not treat such systems as "black boxes", but rather are capable of developing them themselves. In Canada, University of Calgary [10] and University of Quebec [11], both CBSA partners in the past, have a history of developing biometric-enabled behaviour screening technologies. Another major player in this field is the National Research Council of Canada (NRC), the world leader in 3D range data sensing and in semantic and sentiment text analysis. Canada is also traditionally very strong in Artificial Intelligence and Neural Networks research, which contribute to the advancement in the field. Universities of Alberta [12], Toronto [13] and Quebec [14] are among the top-ranking universities worldwide in this area. Collaboration with these institutions is highly encouraged for continuing further advancement of this technology in Canada.

4 AVATAR kiosk at the CBSA

4.1 Four advantages of AVATAR kiosk

The AVATAR kiosk is one of the most known BEBS implementations [3]. Figure 4-1 shows its main components. The main competitive advantages of AVATAR kiosks compared to other work done in automated emotion sensing are the following:

- Realistic animated recreation of the virtual interviewer (see the figure) capable of imitating normal human facial reactions such as blinking and grimacing, which is very important in creating the much desired impression of very powerful and very intelligent machine on a traveller;
- Effective and robust design that automatically adjusts cameras to eye level, which allows capturing of faces and eyes at good quality at all times;
- Highly configurable code for modifying interview scripts;

- Many years of building, tuning and testing the system, including during the current project with the CBSA.

It is noted that the UA team who developed the AVATAR holds two US patents related to the development of IATS / BEBS systems [15,16].

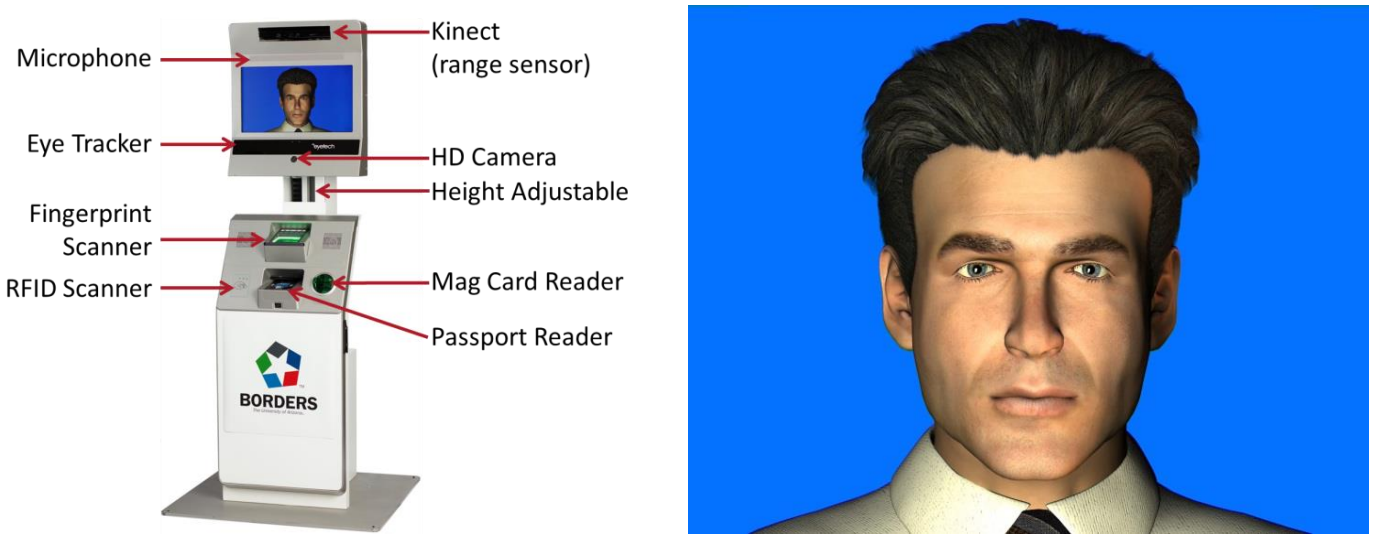


Figure 4-1 AVATAR kiosk design and computer-animated AVATAR virtual agent.

4.2 Test design

In March 2016, through the joint effort of UA and the CBSA, an AVATAR kiosk was brought to the CBSA for a prototype testing exercise. Over eighty volunteers from the CBSA and other government and academic partner organizations have participated in this exercise, making it possible to collect what is now the largest data-set for the research in automated behaviour screening for ABC applications. All personal identifiable information (PII) was removed from the collected data, which were then analyzed using a new evaluation framework described in the next section.

The test was designed to test the AVATAR technology in two behaviour screening problems: detecting "imposters" and detecting "smugglers". Each test participant was issued a plastic machine-readable ID card and was asked to pack a travel bag (see Figure 4-2). Half participants played imposters/smugglers ("Liars"); they were asked to hide a contraband item (shown on the bottom right in Figure 4-3) in their bag and had false information on their ID cards (false first name and false age). The other half of participants played regular travellers ("Non-liars"). A video describing the test has been prepared and can be provided upon request.

Figure 4-2 Plastic ID card issued for a test participant. A participant packs his backpack with travel items.

A two-minute interview script that was prepared for the AVATAR is conceptualized in Figure 4-3. Horizontal axis shows the actions that a traveller had to perform. These actions were the answers to seventeen (17) questions asked by the AVATAR agent during an automated two-minute interview. . Questions where Liars needed to lie are marked red.

Eight questions were supplemented with images shown at bottom: one showing person's ID information and seven showing the images related to custom declaration questions. Users were asked to look at those images to confirm their answers to the questions. Vertical axis shows biometric features (person's facial emotions, voice characteristics, eye tracking information and pupil dynamics) that were obtained using the sensors (video camera, microphone, eye tracker) installed in the kiosk.

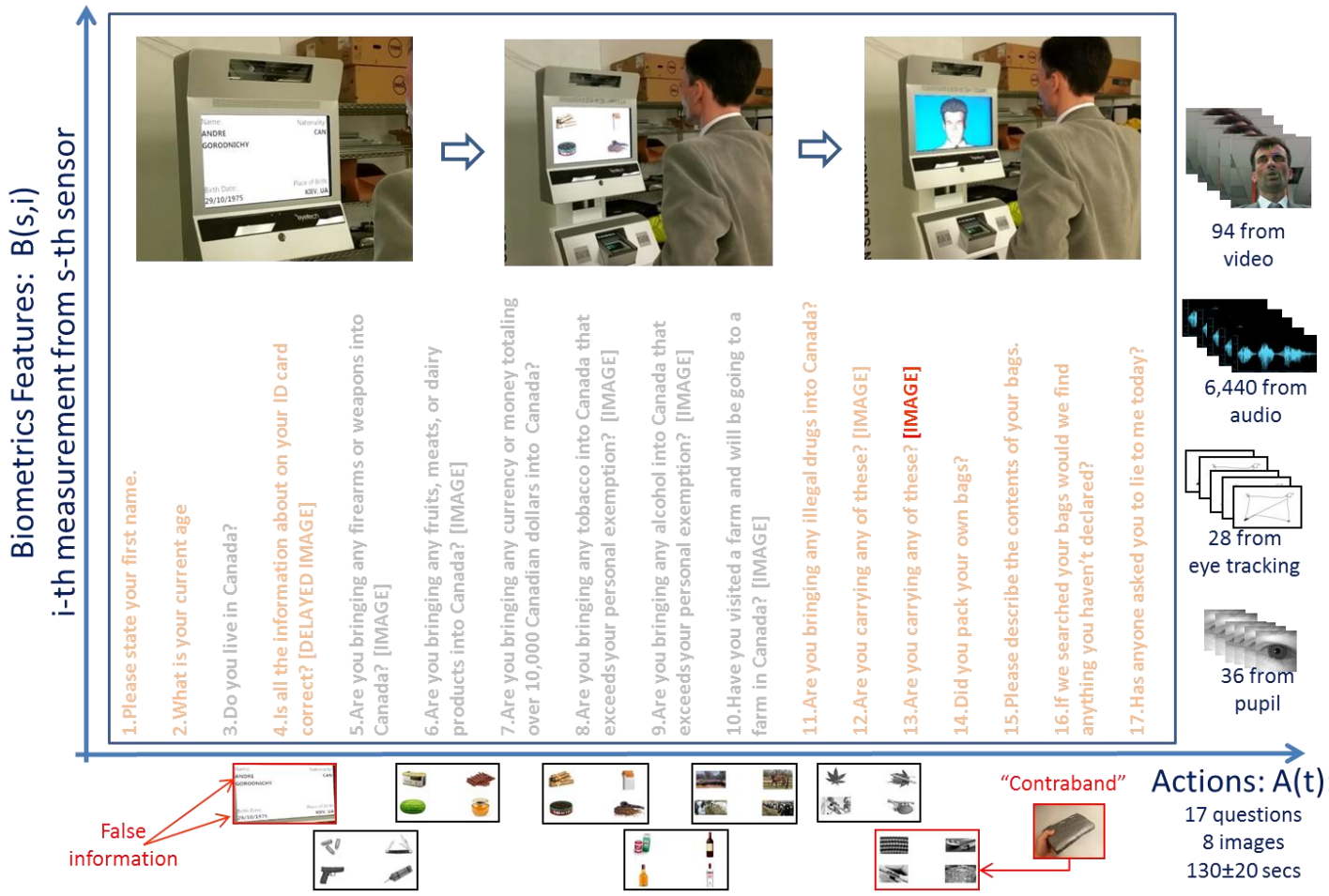


Figure 4-3 Graphical representation of the AVATAR experiment at the CBSA.

Large rectangle in the middle of the figure, within which interview questions and interview process images are shown, represents the total number of all features measured during the interview, the number of which is equal to the number of all measurements from all sensors B(s.j), shown along vertical axis, times the number of all actions (responses to interview questions) sampled over time A(t), shown along the horizontal axis.

4.3 Features measured

Third party software was used to extract features from the data recorded by the sensors. First, raw statistical measurements, as registered directly by the sensor, are obtained. Then these measurements were processed to produce the total of nineteen (19) higher-order statistical measurements (*functionals*) per question, such as mean, standard deviation, range, quartiles, up-level times, regression slope, intercept and others (See Appendix for more details). In such a way, for each action (i.e., each of 17 asked questions, eight of which were supplemented by images), the following numbers of features were computed at each interview:

- From HD (30FPS 720p) video-camera: 5*19*17 facial emotion measurements (Happy, Surprised, Neutral, Disgusted, Sad) computed on features extracted from video using Intraface software [17] (see Figure 6-7);
- From microphone: 6440*17 vocalic measurements which include traditional acoustic measures (f0, response latency, intensity, etc.) extracted from raw audio using Praat software [18] and Interspeech Computational Paralinguistics Challenge acoustic features extracted from raw audio using OpenSmile software [19];

- From eye tracker (pupil and proximity data): 2*19*17 pupil size and distance to camera measurements computed directly from eye tracker data;
- From eye tracker (eye tracking data): 4*7 eye fixation measurements computed from eye tracking data, indicating how long a person looked at each of four quadrants in the seven images.

This resulted in a total of over a million biometric and non-biometric measurements recorded at each interview. As will be shown in the next section, of these measurements, there is only a very small fraction (between 20 and 40 measurements) that is useful. Other measurements present noise, irrelevant or highly correlated (repeated) measurements.

In addition to the above listed measurements, the system has also recorded eye and face tracking data (over 350,000 vertical and horizontal coordinate measurements recorded at 60Hz by the video camera and eye tracker for each interview session). These however were not used for the analysis.

5 Methodology for designing and evaluating BEBS systems

5.1 Limitations of ISO biometric evaluation standards

Current biometric standards, such as those developed by International Standards Organization (ISO) Sub-Committee on Biometrics (SC-37) and presently used by industry and academia [3], are limited to the use of biometric data for identification purposes only (such as 1-to-1 identity verification and 1-to-N identity search). As a result, no formal guidelines have been developed to date for evaluating systems that use biometric data for behaviour screening applications. In order to evaluate BEBS technology, and AVATAR kiosk, in particular, we had to develop new BEBS-related terminology and BEBS evaluation methodology.

In this regard, a list of new biometric terms related to BEBS applications has been developed and submitted to International Standard Organization (ISO), presented in a separate report [5], and a novel framework for designing and evaluating such systems has been developed, further described below.

5.2 BEBS systems vs. traditional biometric systems

The key difference of BEBS systems from traditional biometrics comes from the fact they rely on biometric features that, when used individually, are of very low discriminating power. This is seen in the Appendix, which compares the density of matching scores obtained from voice for “Liars” vs. “Non-liars”. It is seen there that density distributions for opposing classes are only very slightly different from each other. This is in sharp contrast to traditional biometrics, where density distributions of opposing classes are very different and well separated (See [20], for example, where the same voice modality is evaluated for identification applications).

The reason for such low discriminating power of features in BEBS is that they are compared, not to the features of the same person (as done in traditional biometrics), but to some “averaged” features (models) representing their class (Liars). These models are computed based on some historical data and the previous knowledge of deception signals described in Section 3, and they may not be assumed perfect.

To compensate for low discriminating power of features, BEBS systems use large quantity of those features, accumulating them over a period of time (e.g. over two minutes at 20 Hz sampling rate, as done in AVATAR). This is illustrated in Figure 5-4, where the total number of measured features (represented by a large perimeter box in the figure) is very large, equal to the number of all measurements from all sensors B(s.j), shown along vertical axis, times the number of all actions (responses to interview questions) sampled over time A(t).

Such large accumulation of features in BEBS systems further differentiates such systems from traditional biometrics, where recognition is made from only few biometric features and mostly from a single action (such as looking into iris camera). The result of such accumulation is that most of features do not contain information related lying, and those features that are related to lying may not be known in advance.

Based on these two critical differences of BEBS from traditional biometrics, the principles for designing and evaluating BEBS systems are developed.

Figure 5-1 Key principles for building and testing BEBS systems.

At Design stage, best feature-actions are identified (shown as small grey rectangles inside a large rectangle). At Evaluation stage, the decision is made in such as way it is does not generate many false alarms and is easy to interpret by officer, e.g., by showing features that have been found to be abnormal (marked red, and shown on officer's laptop screen).

5.3 Four design principles

The main four principles for designing BEBS systems are defined as follows (see Figure 5-1), further elaborated upon below.

1. *Design objective:* Identifying Actions and Biometric features that provide the best statistically quantifiable discrimination between lie and non-lie behaviours;
2. *Performance metrics:* Using metrics suitable for detection low frequency events, as done in Video Analytics applications;
3. *Criteria for success:* Measuring the success in terms of likelihood to improve the status-quo recognition rates, as done in Clinical Trials, rather how this is done in traditional Biometrics;
4. *Operator-centered design:* Designing recognition models and detection visualization interfaces such that are easily interpretable and efficient for humans.

5.3.1 Dealing with sparsely-spread mostly noisy measurements

The performance of BEBS systems needs to be examined along two dimensions: one related to the biometric features that are measured on a person, and another related to the actions that the person is required to perform while his or her behaviour is being analyzed by a biometric system. In such a way, the biometric features and the actions that elicit the best discriminatory power of the system can be selected and compared to one another.

The total number of all captured features (i.e., the number of measurable actions and measured features per action) during the screening process can be very large – potentially several orders larger than that in conventional biometrics. Evidently not all performed actions and not all measured data are "useful" (i.e., related to a person's stress),. It fact, it could be estimated that most of measured data (>90%) are not "useful". They are, what is called in machine learning parlance, noise. At the same time, there does exist small portion of measured data (<10%) that are indeed "useful" for stress detection. These data however are dispersed over a long range of actions and measurements dimensions, and are not known in advance

In the absence of knowledge about "useful" vs. "non-useful" features, two simplified ways of designing BEBS have been used to date. The first one uses only one, or very few, measurement(s), e.g., only the pupil change at a question about age. The second one uses all recorded measurements, i.e., the entire large rectangle shown in Figure 5-1As presented in the next section, both of these designs are capable of "detecting" liars – in a statistical sense, i.e., that the performance of thus designed system is better than flipping a coin and possibly better than that by humans. However, it is understood these two designs, even though are the easiest to implement, may and should be improved, which is our objective.

Therefore **(Principle 1),** the key objective of the BEBS design is to find these (possibly highly narrow pockets of) useful data (features and actions) that yield best possible recognition (as shown on the left side of Figure 5-1).

5.3.2 Parallels with Video Analytics:

Traditionally biometrics systems are evaluated using detection error tradeoff metrics, which measure *True Positive Rate (TPR)* and *False Positive Rate (FPR)* (or its inverse *True Negative Rate: TNR=1-FPR)*. This is how the accuracy of BEBS systems, including AVATAR, has been traditionally reported. However, for border control applications, where the traffic is very large (e.g., over a million travellers daily), such metrics are inappropriate and should not be used; e.g., reporting FPR=70% at FPR=50% would mean that half million travellers are falsely flagged, which should never be something to be contemplated in light of the Presumption of Innocence principle!

Therefore **(Principle 2),** the BEBS performance should be measured using the metrics that are suitable for low-frequency events, as done in Video Analytics applications (e.g., automated detection of dropped luggage) [5]. There instead of TNR, the ratio of True Alarms to all Detected Alarms, called *Precision*, is measured. The Precision metric can be measured for any size traffic volume and is reported per unit of time. See Table 5-1 for illustration of these concepts, using the data obtained from AVATAR testing.

Table 5-1 Performance accuracy metrics used in Biometrics and Video Analytics.

	All Liars (42)	**All Non-liars (40)**	All alarms	**Metrics for rare event detection**
Believed to be **Liars** (Alarms) after interview By computer vs. By human*	22 vs. 10 **TPR**=22/42 **TPR**=10/42	30 vs. 5 **FNR**=30/40 **FNR**=5/40	52 vs.15	**Precision = TPR / All alarms** 22/52 vs. 10/15
Believed to be **Non-liars** after interview By AVATAR vs. By human*	20 vs. 32 **FPR**=20/42 **FPR**=32/42	10 vs. 35 **TNR**=10/40 **TNR**=35/40	All non-alarms	**Recall (a.k.a Accuracy) = TPR**

***** *For computer, a sample of results from AVATAR testing is used.*
For human, the results are hypothesized based on visual observations of interviewees by the author of this paper.
Numbers are provided for concept illustration purposes only and should not be used as a reference on system/human performance.

5.3.3 Parallels with Clinical Trials:

The most radical way the BEBS systems differ from traditional biometric system is the fact that, they should be expected not work for everyone, regardless how good are the measurements and who does the analysis, a human or a machine, become some people are just much better lying than others. This brings analogy with evaluation of medications, which may or may not work for some patients, depending on their medical history. Hence, the use of statistical metrics, such as confidence rates on probabilities of success, as done in clinical trials, rather than how this is done in traditional biometrics, is more appropriate for quantifying the quality of BEBS recognition.

Therefore **(Principle 3),** the success of the BEBS should be measured in terms of improvement (or lift[2]) over manual screening, rather than in absolute accuracy / precision numbers A technology with any positive improvement (lift larger than 1) can be potentially recommended for deployment, depending on its cost and other benefits or side-effects.

[2] **lift** is the ratio of the correct responses obtained with a new technique and the correct responses obtained by status-quo technique (https://en.wikipedia.org/wiki/Lift_(data_mining))

5.3.4 Auditable and easy to interpret

Finally, as a natural consequence from previous principles and further distinguishing our methodology from status-quo biometrics evaluation/design practices, is the fact that it evaluates the success of BEBS systems in relationship to how well it *assists* (rather than *replaces*) a human officer in detecting a threat .

Therefore **(Principle 4),** the quality of the recommendations made by the system is judged not only on the merit of error rates, but also on the merit of their interpretability by officers, as shown on right side of Figure 5-1. The predictive models that are easier to interpret by end-users may be preferable to better performing predictive models that are harder to interpret.

5.4 The data analysis workflow for designing BEBS

Based on the principles described above, the following three-step process is established for designing BEBS systems. All these steps are performed during the Design and Tuning Stage (left side of Figure 5-1).

5.4.1 Step 1: Data preprocessing

Step 1.1 (Detection of noise and outliers). Our experiments showed that, in contrast to traditional (ISO-defined) use of biometrics, because of lack of constraints on user motion and insufficient time to properly calibrate sensors for each new user, a significant portion of measured data (> 10 %) may contain missing and unrelated measurements (outliers). These data need to detected and removed. Otherwise they may lead to false recognition results.

Step 1.2 (Data calibration. Optional step). Environmental conditions may vary. Additionally, human behaviours may vary substantially (e.g., some may closer to kiosk or talk lauder from the start) . Therefore, some part of the interview can be used to collect user's "baseline" measurements. These "baseline" measurements are then used to calibrate all other measurements, so that to minimize the effect of these variations on the deceit recognition performance.

5.4.2 Step 2: Feature correlation and effect analysis

Step 2.1 (Feature correlation analysis). Because of large quantity of measured features, many of them are correlated. Care should be taken when deciding which features to remove and retain for analysis. Features that come from the same modality should be removed, as they do not provide additional evidence, but rather are the derivatives from the same raw measurement (as most half of features in pupil modality). In doing that the preference should be given to those features that are easiest to interpret by humans. In contrast, all correlated features coming from different modalities (such a proximity and pupil modality) should be retained. Appendix provide examples of such analysis.

Step 2.2 (Feature performance analysis). The features are analyzed for the effect produced on them by lying. Two classes of such effect have been recognized as an outcome of our study. The first class relates to the act of lying (annotated as bLied in our analysis). The second class is related by being a liar (annotated as bLiar), meaning that a person is saying truth for a particular question, however, some parameters of his/her behaviour are already affected because he/she is a Liar. Three orders of feature analysis of increasing complexity are proposed to detect best performing features.

The simplest, *Order 1, analysis* consists in comparing (plotting) the basic statistical metrics (such as average and variation) for each feature to see if (and how) their score varies for different classes. This allows one to visually select the variables (features) and questions that have better discriminatory power. This also allows one to find the features that are affected by *lying* and by *being a liar*. The example of such analysis is shown in Figure 5-2, which plots average scores for a voice feature (dursec) and a pupil feature (pupilmean) for each question; the curves show the smoothed averages and the bars show the actual mean plus/minus variation values, averaged over all users.

Another example of such analysis is shown in Appendix, where matching score densities for all vocalic features are plotted for two opposing classes: bLiar=0 vs. bLiar=1 (first image) and bLied=0 vs. bLied=1 (second image).

The feature correlations for opposing classes are also shown there and an important discovery is made. The act of lying and being a liar effect different features! Features that are effected by lying are mainly located on the right of the feature spectrum, while features that are effected by being a liar are mainly located on the left. This finding may assist both system designers and human analysts. It is also further validated and quantified using higher order analysis.

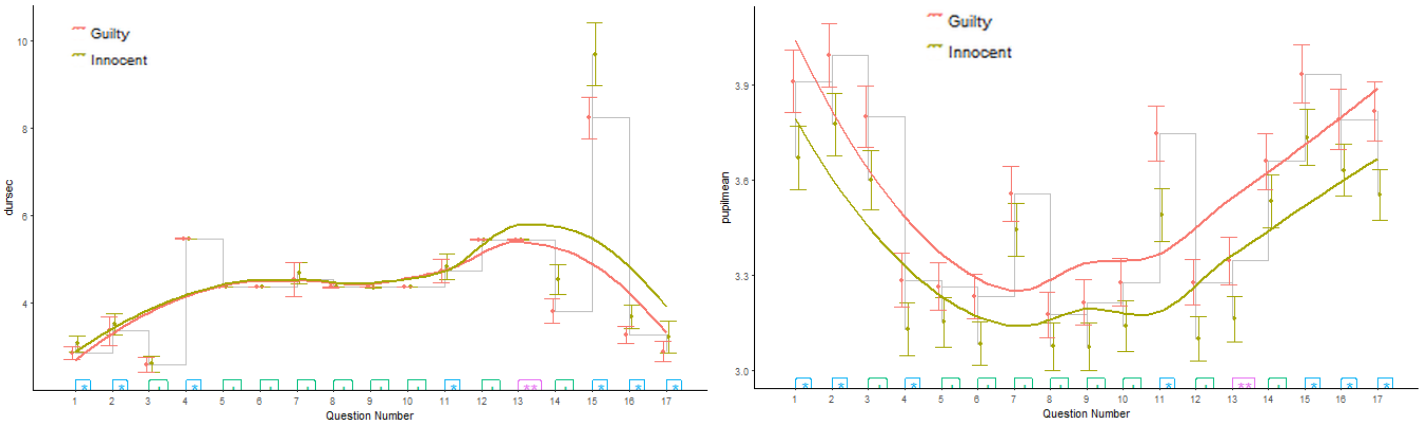


Figure 5-2 The average score values, shown as boxplots (mean +/- variation) and a smoothing curve, for Liars (Guilty) and Non-liars (Innocent) on each question. See Figure 4-3 for the description of questions.

[Visual examination of these scores reveals that some features are affected by lying, e.g., a vocalic feature *dursec*, (shown in left), while others are affected by being a liar, e.g., pupil feature *pupilmean* (shown at right).

Order-2 analysis applies statistical hypothesis two-sample tests, such as T-test or Wilcoxon signed-rank test [20], to measure the probabilities that scores from two opposing classes (liars vs. non-liars) belong to the same statistical distribution. This analysis can be conducted for each feature and each question. The feature-question pairs that have probability P<5% are kept as the best for system design.

Figure 5-3 shows an example of the best five feature-question pairs for pupil modality identified using this technique. The probability of lying to *not* effect these features is shown on the top. Question number is appended to the name of the feature shown on vertical axis. The figure also shows the actual feature scores measured for 82 AVATAR test participants (shown as dots) and the box-plots (mean +/- variation) for each class.

Using this analysis it is found that out of all measured pupil features (large rectangle in Figure 4-3), less than 5% are useful in statistical sense. The same situation is observed for other modalities.

Figure 5-3 Best performing pupil-question features are identified using statistical hypothesis two-sample tests.

Two-sample statistical hypothesis test may not be used when the number of samples is small and when the effect due to inter-personal differences (the so-called random effect) is comparable to the effect of lying (the so called fixed-effect, or effect under investigation) In this case, a more advanced, *Order-3 analysis*, that takes into account both fixed and random effects is applied. It is commonly used in ecology and linguistics and is applied for our work [22].

Briefly, according to [22], two linear models are built using mixed-effect regression (*lmer* function in R [9]), one with effect under investigation and another one without it. For example, to investigate the effect of being a Liar on pupil size, we have:

```
model.null <- lmer(feature_ pupilmean ~ question + (pupilmean || user)
model.withEffect <- lmer(pupilmean_ ~ question * bLiar + (pupilmean || user)
```

The first (NULL) model says that pupil size changes with question, which is the integer number from 1 to 17, for everyone (fixed effect). It also says that everyone may have their own type of change (random effect). That is, regression slope and intercept (i.e., the change gradient, absolute value and starting value) for each person may be different, which is designated using `(pupilmean || user)` notation. The second model is identical to the first one, except that it says that pupil size changes not only with question *but also* the integer value bLiar, which is 1 for Liars and 0 for Non-liars. Then the likelihood ratio (LR) of one model being better than the other is computed by applying analysis of variation (*anova* function in R) with both models and measuring its Chi-Square test value, which presents the probability that models are same. The probability of the effect is computed as one minus Chi-Square test value.

```
result<- anova(model.withEffect,model.null);
ProbabiltyOfEffect <- 1- result$`Pr(>Chisq)`
```

Key outcomes of applying Order-3 analysis are presented in the next section, while Appendix C provides additional details and results related on this powerful type of analysis.

5.4.3 Step 3: Predictive model selection

Step 3.1 (Building an interpretable predictive model based on the selected biometrics features). There are many techniques available in R to build models that predict output value (which is 0 and 1 for Non-Liar and Liar, in our case) from an input vector (which is a vector of selected biometrics measurements, in our case). However, techniques which are easier to interpret, such as linear regression or decision trees, may be preferable for the BEBS application to more complex techniques such as neural networks, which are hard to interpret by humans.

Additionally, different models can be built for different modalities and different questions, and then the best one (or a combination of several good ones) can be selected separately for each modality and/or question to improve the overall recognition performance of the system.

The above steps are typical for the statistical analysis and data science. They are not however yet a part of current biometrics recognition practices, which are simply based on comparisons of a probe image to stored images and limited to producing match/non-match rate statistics only.

The final step is used both for selecting the best model during the design stage and then, once the best model is selected, during the evaluation stage (refer to Figure 5-1):

Step 3.2 (Measuring the performance of the system using the metrics suitable for rare-event detection). As emphasized earlier, it is recommended that Precision-Recall metrics (see Table 5-1) be computed and reported for the BEBS systems, as they better suitable for reporting the performance of low-frequency event detection systems.

Good machine learning practices require that data are always divided into three subsets: one used for training the models, one for validating (comparing) various models and selecting the best-performing among them, and finally one used solely for the testing and reporting, so that the reported results are not biased by knowing in advance what type of data will be used.

Unfortunately, in situations when the amount of data is limited (which is our case) such practices may not be possible, and it is common to use two subsets instead of three: one for training and the other for both validation ***and*** testing (reporting). In such cases, the reported results should be considered as optimistic, meaning that in real-life application the performance of the system will likely be worse than reported on test data.

6 Experimental Results

This section presents the results obtained on the data-set collected during the AVATAR kiosk demonstration at the CBSA in March 2016. First, we present the results obtained by two simplified system designs discussed in Section 5.3.1, namely, when all ("useful" and "non-useful") features and questions are used (i.e., the entire large rectangle shown in Figure 5-4) and when only one feature is used. These were original designs used by the AVATAR developers prior to our study. These results are referred to as baseline results.

Then we present results related to further improvement of BEBS accuracy, as per the methodology developed by this study. They are aimed at identifying features that contribute to better performance (so called "useful" features) versus those that do not. These results should assist both system designers (in and human analyst dealing with deceit detection.

Finally, we present other insights learnt from working with AVATAR which should assist designing AVATAR-like technology in the future.

6.1 Baseline results

Below we present the results obtained without applying system design methodology proposed by this report. All features from all questions are fed into machine learning modules. No distinction is made between features that are affected by lying and those that are not.

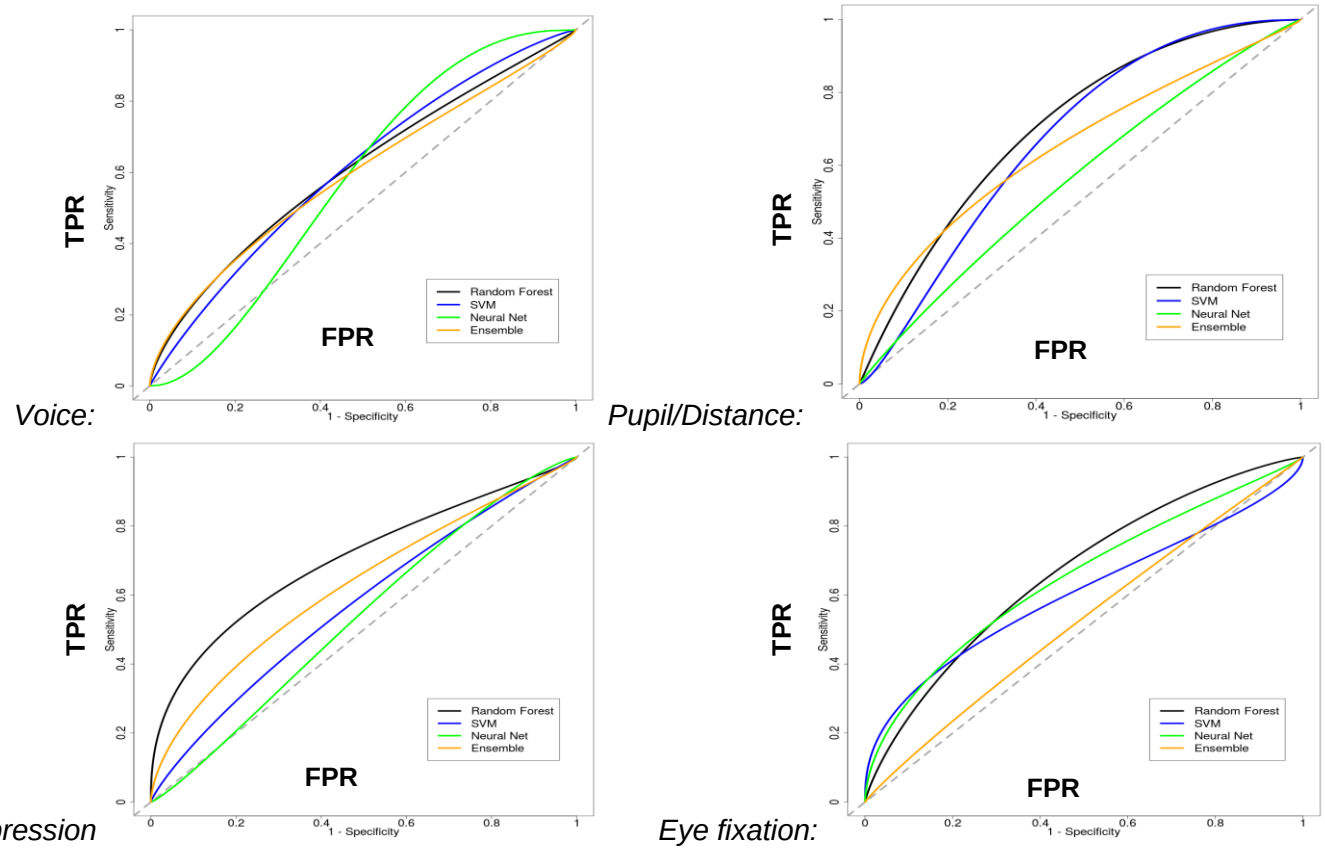


Figure 6-1 Accuracy curves, showing True Positive Rate (TPR) as a function of False Positive Rate (FPR), computed for each modality using all-feature approach.

According to the BEBS evaluation principles described above (Section 5.1.2), False Positive / Negative Rates may not be used for rare-event detection systems. Therefore, the final results for the baseline performance of AVATAR kiosk are presented using the Precision-Recall curves and metrics (refer to (Section 5.1.2 and Table 5-1). Figure 6-2 shows the Precision-Recall curves when all modalities are used. Modality-specific curves are

shown in Appendix F. Based on these curves, the baseline performance scores for AVATAR, for each modality and combined, are presented in Table 6-1.

For comparison, the table also provides Accuracy (a.k.a Recall) results for manual lie detection by humans, which is estimated based on the information provided by UA scientists for the conditions that are similar to those used in the AVATAR test, i.e., when recognizing a lie from a two-minute interview using 17 custom-declaration question. The table also lists advantages and disadvantages of each modality and manual lie detection.

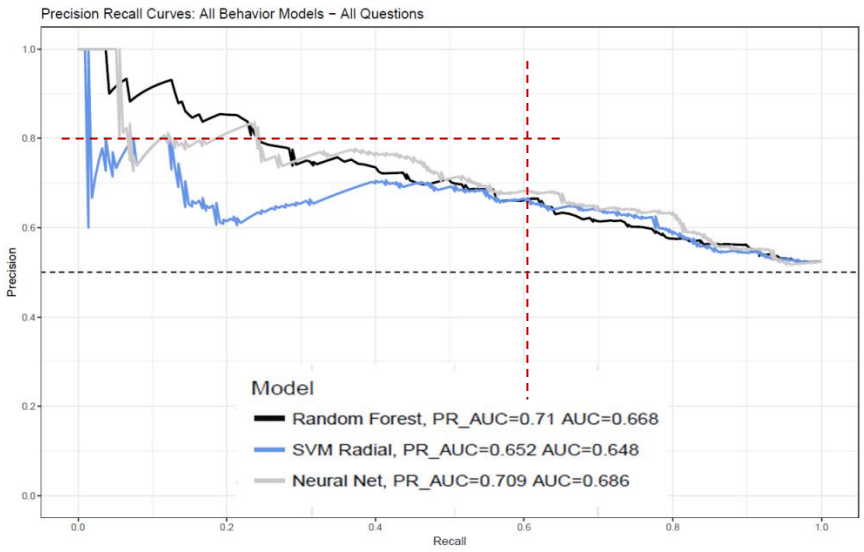


Figure 6-2 Baseline AVATAR performance, using all-questions approach: Accuracy (a.k.a. Recall) is shown as function of Precision, computed using all modalities with three different machine techniques.

Table 6-1 Baseline AVATAR performance, limits and constraints

	With Microphone: Vocalics	With Eye Tracker: Pupil & distance to camera	With Eye Tracker : Eye Fixation	With HD camera: Face emotions	All combined	Humans*
Recall at 60 % Precision**	55-78%	54-65%	60-65%	48-58%	65-68%	45-65%
Dis-advantages	Privacy issues. Good acoustics. No other voices.	Eyes need to be visible at all times	Calibration required	Lowest performance	Needs technical expertise & support	Can't monitor many features at a time. Hard to audit and scale.
Advantages	High performance	No PII recorded. Calibration not required. High performance	No PII recorded	Easiest to capture. Calibration not required.	Scalable, auditable, bias-free	Does not require procuring and maintaining equipment

* *Estimated for the same conditions (using two- minute interview and 17 custom declaration questions).*
** *Based on testing the AVATAR kiosk at the CBSA with 82 participants. See Appendix for the curves from which the numbers are derived.*

Result: It is seen that at Recall of 60% (red vertical line), the Precision of up to 75% was achievable; or, at Precision of 60% (red horizontal line), the Recall in lie detection of up to 80% was achievable. The constructed predictive models are not easily interpretable by humans. The features and questions that contribute to recognition of lies is not known. This is investigated next.

6.2 Questions-specific analysis

6.2.1 Few-features approach: interpretable models

Another approach that was used for reporting the AVATAR performance was that based on building simpler and interpretable models. The best obtained combined result (i.e., based on the model that was found to be the best of all modalities), is shown in Figure 6-3. There the Decision Tree is based on two voice features only : one at question #2 (about age) and the other at question # 16 ("would we find anything you have not declared?"). Appendix E shows the results computed separately for all modalities.

Result: This simple model achieves True Positive Rate of 56 % (32%+ 24%, shown in the blue boxes at the bottom), at the cost of True Negative Rate of 44 % (shown in the green box, at the bottom of the decision tree). The best decision trees computed separately for each modality are shown in Appendix. It is clear from these models that not all features and questions contribute to lie detection. This is analyzed next.

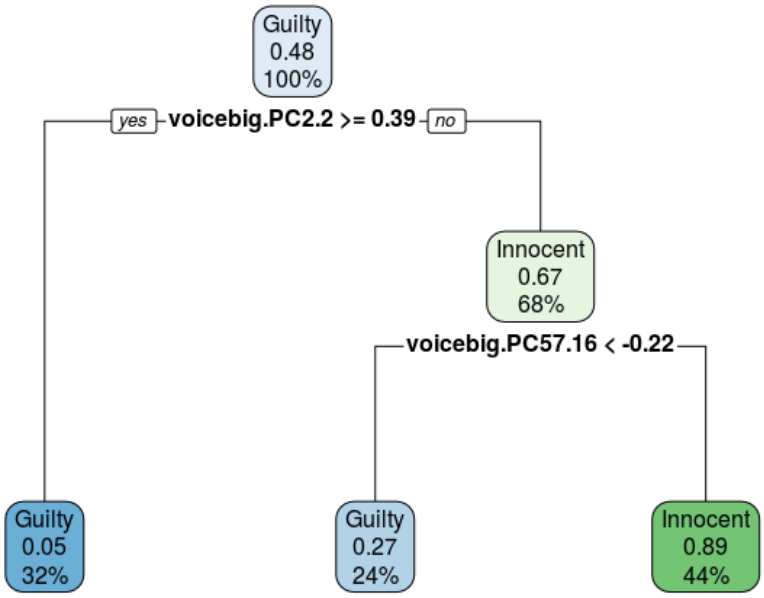


Figure 6-3 Interpretable models: using linear regression and decision trees. Features that are used to make the decisions are shown .

6.2.2 Accuracy by question

Figure 6-4 provides a closer look on the positive and negative contribution of each question to lie detection. The figure shows the combined results that are obtained by applying a single predictive model called Random Forest (which was found to work the best) with all modalities and all features within each modality. The results obtained separately on each modality are provided in Appendix. True Positive Rate (TPR), True Negative Rate (TNR) and combined Accuracy rate (Tcombined), which is the average between TPR and FPR, are shown. Appendix D shows the results computed separately for each modality.

Result: It is seen that some questions help detecting lies more than others, with several of them having TPR > 60%. It is seen that the best- performing questions were those related to lying about age and contraband. Critically however, it is also found that besides guilty-knowledge-related questions, other questions may also contribute to deceit detection, and that some questions may likely lead to wrong prediction, predicting with accuracy less than 50%, i.e., worse than flipping a coin.

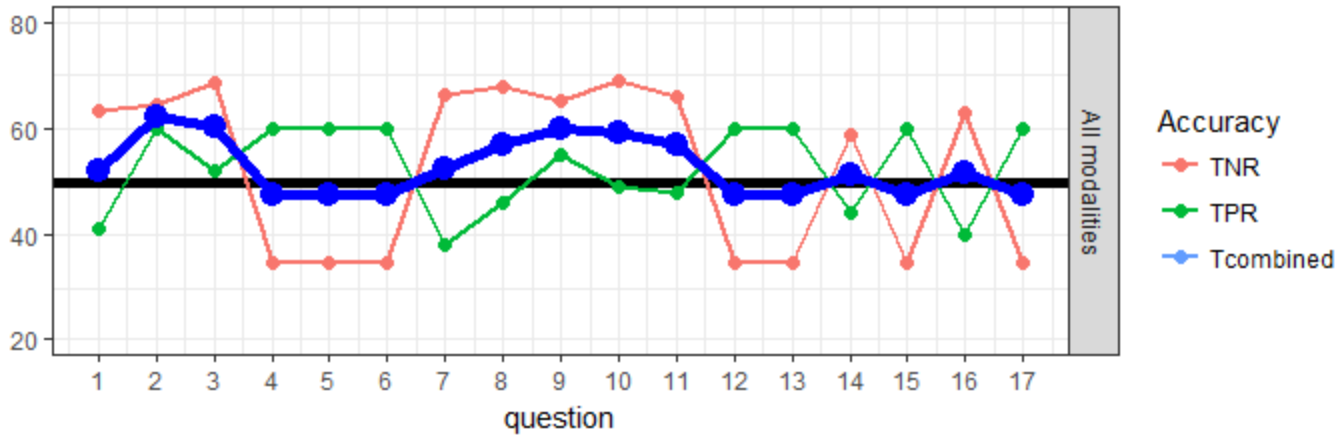


Figure 6-4 Accuracy by question combined over all modalities.

6.3 Best and worst deception indicators

This subsection presents the results related to further improvement of deceit detection technology by means of identifying which behavioural patterns are related to lying and which are not. The mixed-effect linear regression *Order-3 analysis* (Section 5.2.3) was applied for all features in each modality, keeping the person's id as a random effect and having Liar and Lying variables as fixed effects.

Entire results of these experiments are presented in [28], some are also shared in Appendix, while Figure 6-5 provides just one highlight of these results, related to voice features. The figure shows the result of applying *Order-3* (mixed-effects) analyses and a simpler *Order-2* two-sample test analysis. The bars shows the probability of the causal relationship between the feature values and being a liar (shown on the left) and lying (shown on the right) computed using two types of analysis.

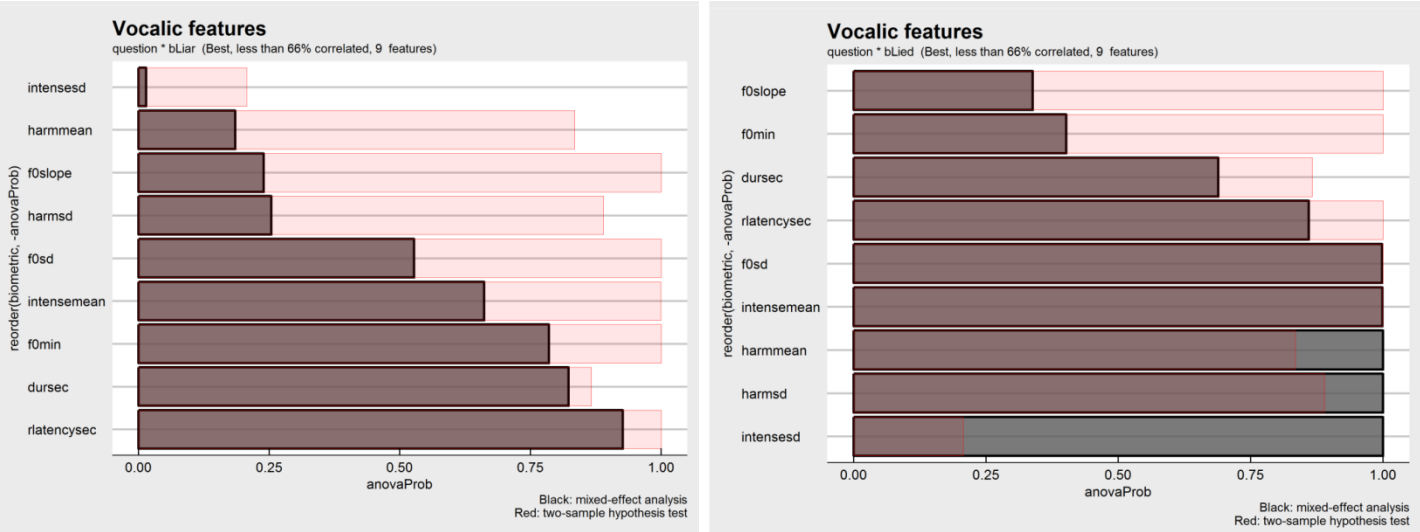


Figure 6-5. Best and worst deceit indicators in voice modality, computed using mixed-effects regression (dark bars) and two-sample statistical hypothesis tests (red bar) for the effect of being a liar (left) and lying (right).

Further to this, Table 6-2 puts together the findings related to the identified best performing feature for each modality. The results in the table should be read as follows: *"Being a liar effected the value of [Modality feature] with probability [Probability of change], changing its average value by over [Change*] percent".

Table 6-2 Mixed-effect analysis results: probabilities and the amount of change due to being a liar.

Modality feature	*Probability of change*	*Average for non-Liars*	*Average for Liars*	*Change*
pupilmedian (pupil size)	P=0.9991	3.3583	3.5479	>5% ↑
distancemedian (distance to camera)	P=0.9853	71.3045	69.9066	>2% ↓
f0max (voice loudness)	P=0.8447	171.7472	187.1196	>5% ↑
happymedian (facial expression)	P=0.9094	0.2533	0.0940	>50%↓

Most discriminating feature in each modality is shown.

Results: The obtained results confirmed the conjecture expressed earlier that lying and being a liar affect different kinds of features, and provided the qualitative metrics, on the value of all features. It is also seen that mixed-effect regression approach provides more meaningful results than traditional two-sample hypothesis test which shows almost all features as "useful". The modality specific findings are listed below.

For example for voice modality (see Figure 6-5), feature *intensesd* (which signifies the variation invoice intensity while responding to a question) was found to be the most powerful for detecting lying (with P>0.9). It is however a very poor discriminator when a person does not lie (P < 0.1). At the same time, it is the opposite voice features *dursec* (which signifies the delay between the question and response) and *f0min* (which is the primary voice frequency), which are found to be good for flagging liars (P=0.85 and P=0.80), but worse for detecting the actual moment of lying (P=0.7 and P=0.32).

Similarly in pupil dynamics modality (see in Appendix C), *pupilrange* (which is the difference between the minimum and maximum pupil size observed during a question) is seen highly valuable for detecting the moment of lying (P>0.9), but not at all when a liar does not lie (P < 0.1), in which case *pupilskew* (which describes the shape of the dynamics in pupil change during the response) is found most useful (P>0.95)

Finally, it is also seen that (see Appendix C), when analyzed using mixed effect approach, the proximity features (which characterize the dynamics of the distance between the person and the kiosk) have not been found indicative of lying (all them has low probability in relationship to both lying and being a liar). This disproves the earlier belief that such distance matters.

All this knowledge was not available prior to this study. It expected that it may help now both BEBS system designers and human interrogators to do their job better.

6.4 Additional insights

6.4.1 On use of eye tracking data

An example of eye tracking and fixation data (overlaid over AVATAR screen images) are shown in Figure 6-6. The complexity of eye motion patterns can be seen. Within the duration of the project, it has not been possible to obtain a good algorithm to convert these data into useful features for use in automated lie detection, except for simple counting of time spent in each of four quadrants. At the same time, some evidence has been found (see inset in the figure) that a liar may indeed, consciously or subconsciously, fixate on a target (i.e. guilty knowledge) image. One of 42 liars showed a very prominent fixation on the image showing the drug that she was carrying, whereas no-one of non-liars showed any prominent fixation on any image.

While automated analysis of eye tracking may not be feasible yet, the study has shown that eye tracking data can be made easily interpretable by humans using intelligent Visual Analytic interfaces (e.g. as shown). Such visualization of features may allow the officers to see deception signals that are not visible otherwise by naked eye. For example if a person, for whatever reason, was attracted by a particular image on a screen and that his/her pupil dilated at the same time, an officer may be able to use this information to ask additional questions.

Another observation is that eye tracking data for one fifth of Liars (12 out of 42) appeared to be badly calibrated, meaning unusable for any analysis, and removed from analysis. This is in high contrast to Non-liars all of whom

had well-calibrated eye tracking data. Possibly this insight can be used in future work to improve the discrimination between liars and non-liars.

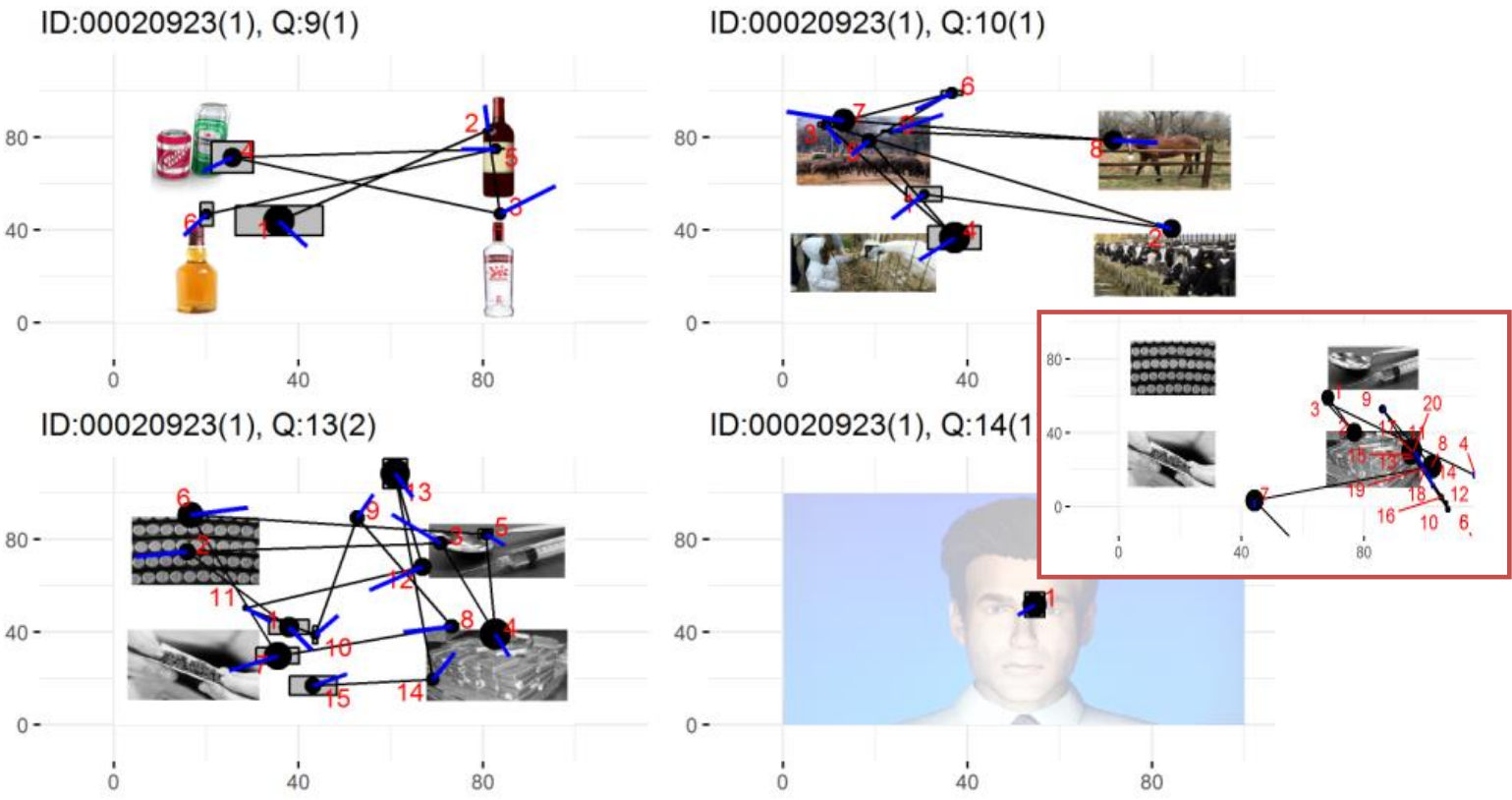


Figure 6-6 Eye tracking results and appropriate visual interface.

Visualizing eye tracking data (trajectory and dwell time) allows one to see where and how long a traveller was looking at the screen in each question. Inset shows a person fixated on an image with the contraband. The numbers in brackets indicate person type (Liar vs. Non-Liar) and question type (for saying truth vs. for lying).

6.4.2 On the use of facial cameras

Typical face tracking and emotion recognition results, which one may expect to obtain from a HD video, are shown in Figure 6-7. The quality of "emotion measurements" computed by the third party program can be seen. The Precision-Recall curves above for facial emotion "measurements" (shown in Appendix) do not show performance much higher than flipping a coin. However, the results from statistical tests and mixed effect analysis shown above did confirm that certain facial expression (such as "happiness") were much more prevalent among non-liars than among liars.

At the same time, it is also found that tracking results obtained from HD video camera are comparable to that of commercial eye trackers and they do not require calibration. Taking into account that cameras are already present in many kiosks and that some kiosks are already designed to align these cameras with eye-level, this makes HD video camera very attractive alternative to eye trackers which, as noted above, suffer from calibration issues.

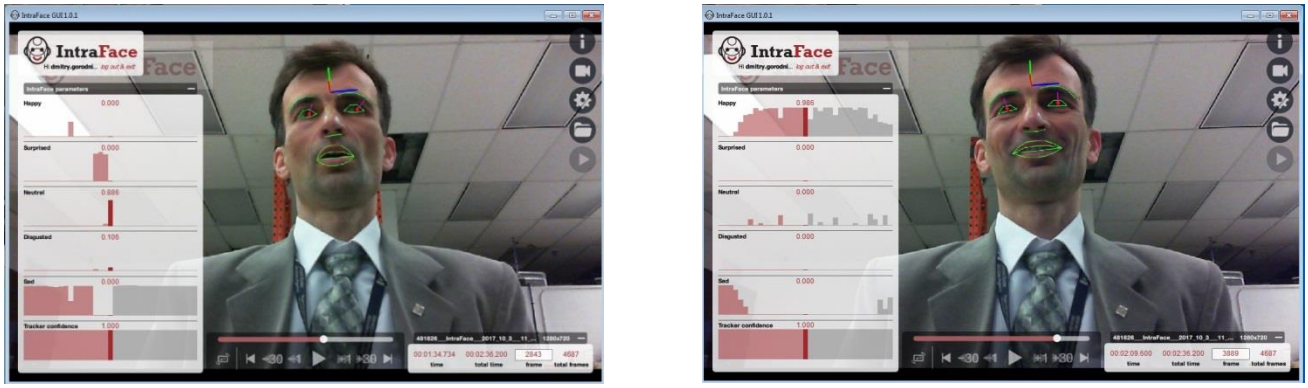


Figure 6-7 Face tracking and facial emotions measurements extracted from a video-camera using Intraface software. Images from two moments during the interview are shown.

6.4.3 On the use of iris camera

An effort was applied to find out if iris cameras such as used for identification can be also used to extract eye tracking and pupil dilation information. It appeared that at the moment, these camera are not be able to do that. The reason is that they are traditionally not designed for continuous streaming of data, but use light strobing to capture only a few images of an eye. This situation however may change in future.

6.4.4 On importance of animated agent

The presence of animated agent has been mentioned earlier as an important factor in creating an impression of an all-mighty agent who can read one's hidden intentions. In this work we highlight yet its other advantage – that of contributing to better calibration and recognition of data. The research on visual attention and saliency [23] tells us that a blinking face will subconsciously attract and control the direction of person's view. This can help to calibrate eye tracking and have less spurious and less spread eye t and face tracking data, which inevitably will contribute to less outliers and better overall performance of the system. More research to quantify the significance of this factor is required.

6.4.5 Other sensors and features considered

Many other sensors and features have been considered for using with AVATAR, however not tested due to project duration limitations. This included using a gyro data on portable tablets that travellers may hold in their hands, weight fluctuation data coming weight sensors such as those already used in some e-gate systems, and finally, the use of blinking patterns, which can be extracted from video as done in [24].

7 Conclusions

7.1 Limitations of the study

The results obtained from the AVATAR test at the CBSA appear better than that by humans (Table 6-1,Table 8-2, Figure 6-2, Figure 8-5, Figure 8-6). However the following limitations of these results need to be highlighted:

1. They are obtained on a very small population size (82 volunteers who participated in the experiment). This is a very small size to be able to extrapolate them to any large size population with any statistical significance.
2. These results are obtained, after several months of tuning the algorithms, following several algorithm tuning iterations. Only the best performing algorithms are shown in this report. In real deployment, once the system is deployed, it may not be tuned.
3. They are obtained on the same dataset that was used to tune the algorithm. As mentioned in Section 5, this creates an "optimistic" bias in the reported accuracies.
4. Volunteers had no advance knowledge of the system. They were not allowed to see the kiosk prior to their interview with it. In real life, this may not be expected from people who come prepared to lie.
5. Furthermore, volunteers were actors, rather than real smugglers / impersonators. Their motivation and skills in lying is questionable. Results of this experiment may not represent real world results.
6. Volunteers were also guided by a study team member, who helped them to start the interaction with the kiosk and who was by their side at all times to resolve any technical difficulties related to operating the kiosk.
7. Results are obtained for this particular scenario and setup that was developed for this test. This scenario involved the same actions for all non-compliant travellers (hiding the same contraband item and falsifying their name and age). They may be quite different for another scenario or setup.
8. Finally, the study did not consider any CELP (Cultural, Ethical, Legal, Privacy) issues related to the use of this technology. These may potentially impose additional constraints on the choice of sensors and data collected, further limiting the accuracy of system performance

Despite a seemingly large number of limitations listed above, this study has identified several ways to further improve the performance of BEBS systems, and make it less dependent on a particular scenario or setup. With another test (or pilot in the field) it should be possible to further improve the accuracy results, using the methodology and new insights presented in this report.

At the same time, the study has also highlighted the fact that simply adding sensors or questions does not necessarily improve the recognition results, which is a common "the more, the better" belief. It can make it worse, if features are not properly tested and filtered

7.2 Technology readiness and limitations

Using the semaphore-like PROVE-IT technology readiness assessment methodology developed in previous projects [6], the Technology Level Readiness (TRL) of the BEBS systems is accessed as being "yellow", i.e., ready for piloting in the field.

Compared to human performance, BEBS technology is shown to offer an increase in the likelihood of detecting deceit. Even though the observed increase may not appear large, it does make this technology very attractive for applications, such as ABC, where a large number of subjects need to be screened and where human factors such as fatigue and bias need to be minimized.

One may expect these results to improve with time as more sensors and biometric features are added. However, as this study demonstrated, it is critical to remember that if not properly designed and tested, adding new sensors and biometric features, may also lead to worsening of the system performance. This is why the development of good standards and building scientific expertise in the area of automated biometric-enabled behavior screening by industry and government stakeholders will be critical for the further development of these systems.

To recapitulate, high scientific integrity will need to be exercised prior to deployment of such systems in the field, because of the elevated risk of falsely flagging travellers due to technology limitations and the absence of public standards for the evaluation of such technology. The principles of ethically designed AI systems, which are being currently actively discussed within IEEE community [25], will need to be followed when designing and deploying BEBS systems.

7.3 Other considerations

It has been mentioned in the report that automated screening systems offer benefits beyond just better Precision / Accuracy rates. They contribute to better transparency, scalability and integrity of ABC decisions. The kiosks that are already used for ABC can do more than what they are doing now. With little modification to their hardware, they may automatically extract the information related to the credibility of people in front of them, thus assisting border officers to make better decisions.

Offering an exciting and pleasant travel experience for people entering the country and being seen by public as a champion driving the technological advances to serve better its clients is another factor not to be discounted, when considering the advantages of AVATAR-like systems.

However there is yet another important factor which we would like to mention in the conclusion of this study. It has to do with the psychological power of people. Using parallels with clinical trials made earlier, it can be called a placebo effect. Just like with medication, when people believe that the technology works, it may work better for them. That is, just by seeing a new and powerful system in front of them, travellers will likely behave in a different and more pronounced way, which may allow the machine and human officers standing by to detect something, which otherwise may not be detected. Extrapolated to large traffic of travellers processed by the Agency, even a very small increase in probability of catching someone will result in dozens of additionally caught smugglers and terrorists.

Table 7-1 Applications potentially suitable for BEBS.

	Application constraints	Sensors that can be used
Frontline applications		
1. Standalone automated (as within PIK)	< 3 mins, No audio Operated by travellers	Video-camera + eye-tracker in kiosk
2. Semi-automated (as a tool in PIL booth)	No eye tracking, Operated by BSO	microphone + video-camera in booth
3. At Secondary examination	None	Video-camera + microphone + eye-tracker
4. Remote ports	VoIP audio quality telecom video quality	Microphone + eye-tracker
Inland applications		
5. Interviews for trusted traveller applications	None	All combined
6. Interview for officer recruitment	None	All combined
7. Self-reporting over the phone	Audio only Lower quality	Telephone microphone

7.4 Recommendations for next steps

For the agency, the following possibilities for further testing of the technology are seen (Table 7-1).

The main opportunity is seen in using the no-voice IATS / BEBS as part of the next-generation Primary Inspection Kiosks (PIK). Kiosks, such as those shown in Table 7-1, are already capable to automatically align cameras with traveller's face, which is the main condition for automated behavior screening. This means that these kiosks may potentially be programmed to perform traveller screening based on captured facial data. The accuracy in deceit detection by such kiosks can be expected to be at least comparable to that of humans. If an eye tracking sensor is installed in such kiosks, then deceit detection accuracy is expected to be higher than that of humans.

IATS / BEBS can also be used as an additional tool for manual screening at PIL booths, where vocal and facial signals can be automatically detected using camera and microphone installed in the booth and shown to the officer for their information.

Another good opportunity for the use of IATS / BEBS is pre-screening (for trusted travellers applications and staff recruitment) and post-screening (i.e., in Secondary examination), where voice recorders can be used and where longer interviews are permissible.

IATS / BEBS can also be used for remote screening (as in self-reporting over a telephone or as at remote unmanned ports of entry) may also be feasible, provided that the audio and video signals are of sufficient quality.

It is important to mention that in all of these applications, IATS / BEBS may be used in both automated and semi-automated fashion. In the first case, the signals detected by sensors are interpreted by the machine. In the second case they are interpreted by humans.

To conclude, IATS / BEBS technology is shown to be ready for testing in the field, with remaining challenges being purely technical - that of properly tuning and evaluating the technology to be deployed. By establishing the methodology for designing and evaluating biometric-enabled behaviour screening technology, this project helps to address this challenge.

References

1. Dmitry O. Gorodnichy, Analysis of Risks and Trends in Automated Border Control: Final Report, DRDC-RDDC-2016-C324. http://cradpdf.drdc-rddc.gc.ca/PDFS/unc256/p804885_A1b.pdf, Executive Summary, Technical Report DRDC-RDDC-2016-C143D. http://cradpdf.drdc-rddc.gc.ca/PDFS/unc229/p803869_A1b.pdf
2. D. Bissessar, E. Choy, D. Gorodnichy, T. Mungham, “Face Recognition and Event Detection in Video: An Overview of PROVE-IT Projects” , Technical Report DRDC-RDDC-2014-C167, http://cradpdf.drdc-rddc.gc.ca/PDFS/unc157/p800402_A1b.pdf
3. AVATAR: The Interrogation Bot, Wired, February 2013.
4. ISO/IEC 2382-37:2012, Information Technology “Vocabulary: Part 37: Biometrics.” Free copies at http://standards.iso.org/ittf/PubliclyAvailableStandards/index.html
5. Dmitry Gorodnichy, New age glossary of biometrics terms for automated border control and video surveillance applications, CBSA Border Technology, Division Report 2016-05, January 2016, Submitted to DRDC June 2017.
6. Dmitry O. Gorodnichy, Diego Macrini, Robert Laganiere, “Video analytics evaluation: survey of datasets, performance metrics and approaches ", Technical Report DRDC-RDDC-2014-C248. Online: http://cradpdf.drdc-rddc.gc.ca/PDFS/unc198/p800521_A1b.pdf
7. Technology Readiness Assessment Results, Video Surveillance Technology Evaluation and Research (ViSTER) portal: https://sites.google.com/site/vistercanada/trl-assessment
8. Deception detection, American Physiological Association, March 2016, Vol 47. No 3. www.apa.org/monitor/2016/03/deception.aspx
9. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/..
10. Biometric Technologies Lab | University of Calgary , www.ucalgary.ca/btlab
11. Université du Quebec, ÉTS : Technologies biométriques , https://www.etsmtl.ca/nouvelles/2011/Technologies-biometriques
12. Artificial intelligence Lab at the University of Alberta: https://www.ualberta.ca/computing-science/research/research-areas/artificial-intelligence
13. University of Toronto, Neural Networks for Machine Learning: https://www.coursera.org/learn/neural-networks
14. MILA : Montreal Institute for Learning Algorithms : https://mila.quebec/mila/
15. Jay F. Nunamaker, JR., Judee K. Burgoon, Aaron C. Elkins, Mark W. Patton, Douglas C. Derrick, Kevin C. Moffitt, Embedded Conversational Agent-Based Kiosk for Automated Interviewing, Filed: January 30, 2013, Publication date: October 10, 2013
16. Nathan W. Twyman, Jay F. Nunamaker, Automated Scientifically Controlled Screening Systems (ASCSS), Filed: May 27, 2015, Publication date: May 4, 2017
17. IntraFace, free research software for facial image analysis: www.humansensing.cs.cmu.edu/intraface
18. Praat: doing phonetics by computer: http://www.fon.hum.uva.nl/praat/
19. The openSMILE feature extraction tool. The Munich Versatile and Fast Open-Source Audio Feature Extractor: http://audeering.com/technology/opensmile/.
20. Dmitry O. Gorodnichy, Michael Thieme, Dave Bissessar, Jessica Chung, Elan Dubrofsky, Jonathon Lee, "CBET evaluation of voice biometrics . In Proc. SPIE Defence, Security & Sensing Conference, 2011. https://www.researchgate.net/publication/228408767_C-BET_evaluation_of_voice_biometrics
21. Wilcoxon signed-rank test, https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test
22. B. Bolker,, et al (2009). Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution, 24(3), 127---135. See more in Appendix
23. Laurent Itti,"Biological Models of Vision and Attention for Face Detection in Natural Scenes. Surprise Approach" , IEEE Workshop on Face Processing in Video, June 28, 2004, Washington, D.C., USA www.visioninterface.net/fpiv04/papers.html
24. Dmitry O. Gorodnichy, Towards automatic retrieval of blink-based lexicon for persons suffered from brain-stem injury using video cameras, IEEE Workshop on Face Processing in Video, June 28, 2004, USA www.visioninterface.net/fpiv04/papers.html
25. Ethically Aligned Design, A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2: IEEE http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

8 Appendices

Appendix A. Description of measured features

During each interview question, which lasts from 5 to 20 seconds, a sequence of biometrics and non-biometric measurements {X}=X1,…Xn is generated by a sensor for each modality according to the sensor sampling rate (eg.60 Hz for eye tracker). This sequence can range from a dozen to over a thousand numbers and is used to generate 19 high-order features (functionals) for each question that are used for recognizing the deceit. These features are the following:

1.	"[X]mean"	Mean
2.	"[X]median"	Median
3.	"[X]sd"	Standard Deviation
4.	"[X]max"	
5.	"[X]min"	
6.	"[X]range"	Range
7.	"[X]skew"	skew
8.	"[X]kurtosis"	kurtosis
9.	"[X]quartile1"	quartiles 1 & 3
10.	"[X]quartile3"	
11.	"[X]iqr"	inter-quartile range (IQR)
12.	"[X]percentile01"	1% & 99% percentiles
13.	"[X]percentile99"	
14.	"[X]percentilerange"	percentile range
15.	"[X]upleveltime75"	up-level times – percentage of time the value was > 75% of its range
16.	"[X]upleveltime90"	
17.	"[X]regslope"	Regression Slope,
18.	"[X]reginterсept"	Regression Intercept
19.	"[X]regrmse"	Regression root mean square error (RMSE)

where [X] is the name of the measured modality, such as :

- pupil, distance – measured by eye tracker
- f0, harm, intense – computed by Praat, and >100 features computed by OpenSmile from audio
- happy, surprised, neutral, disgusted, sad – computed by IntraFace software from video

Within each modality, high-order features can be highly correlated. In this case, these features need to be removed, as they do not provide additional evidence, but are simply repetitions of the same knowledge. Among different modalities, high-order features may also be correlated. In this case however, they should not be removed from consideration, because they do add additional evidence towards the recognition objective. Figure below shows inter-modality correlations and feature names.

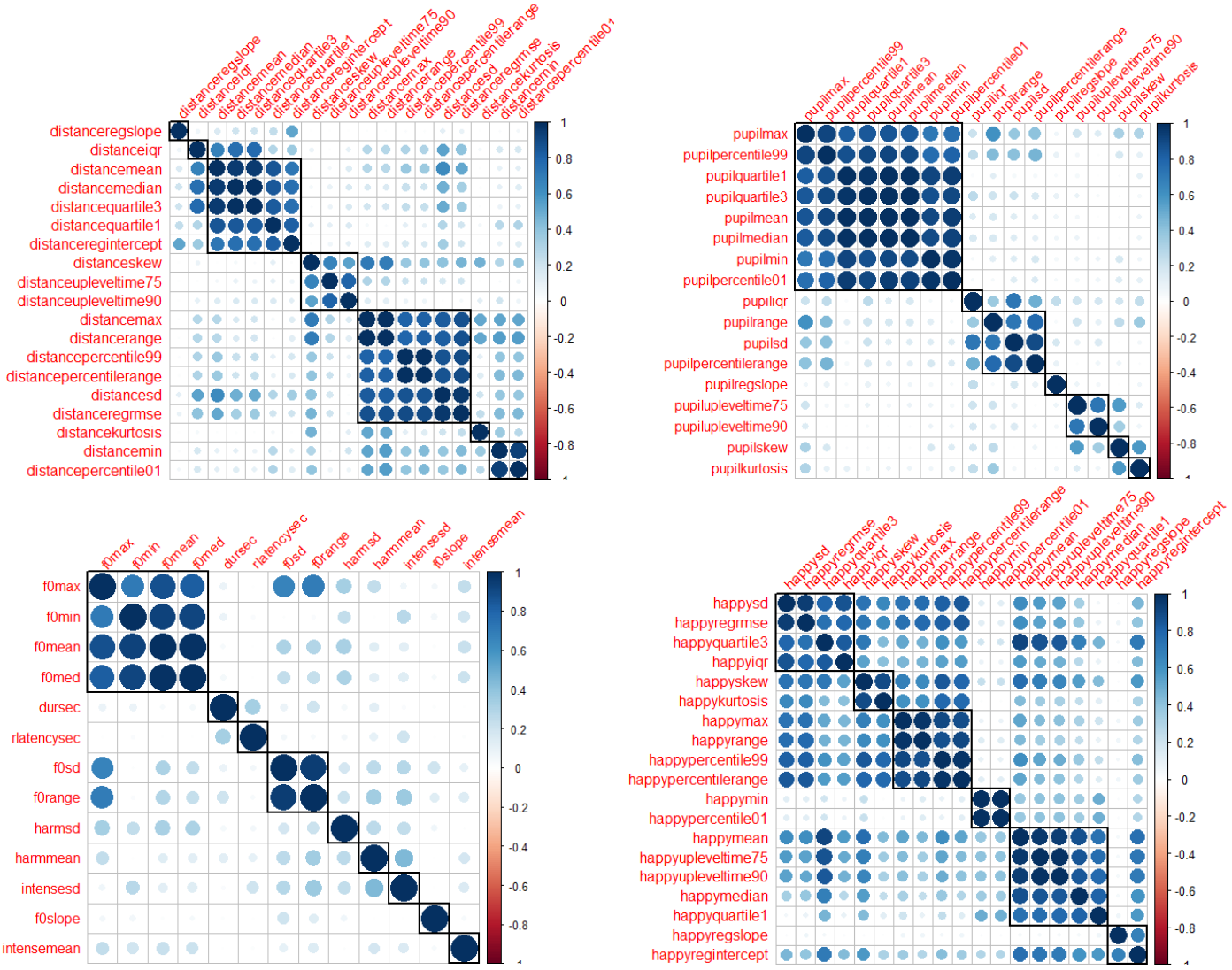


Figure 8-1 Correlation of pupil/distance features obtained from eye tracker (a) and vocalic features obtained from audio using Pratt software (b).

Appendix B. Visual analysis of the effect of lying and being a liar on voice features

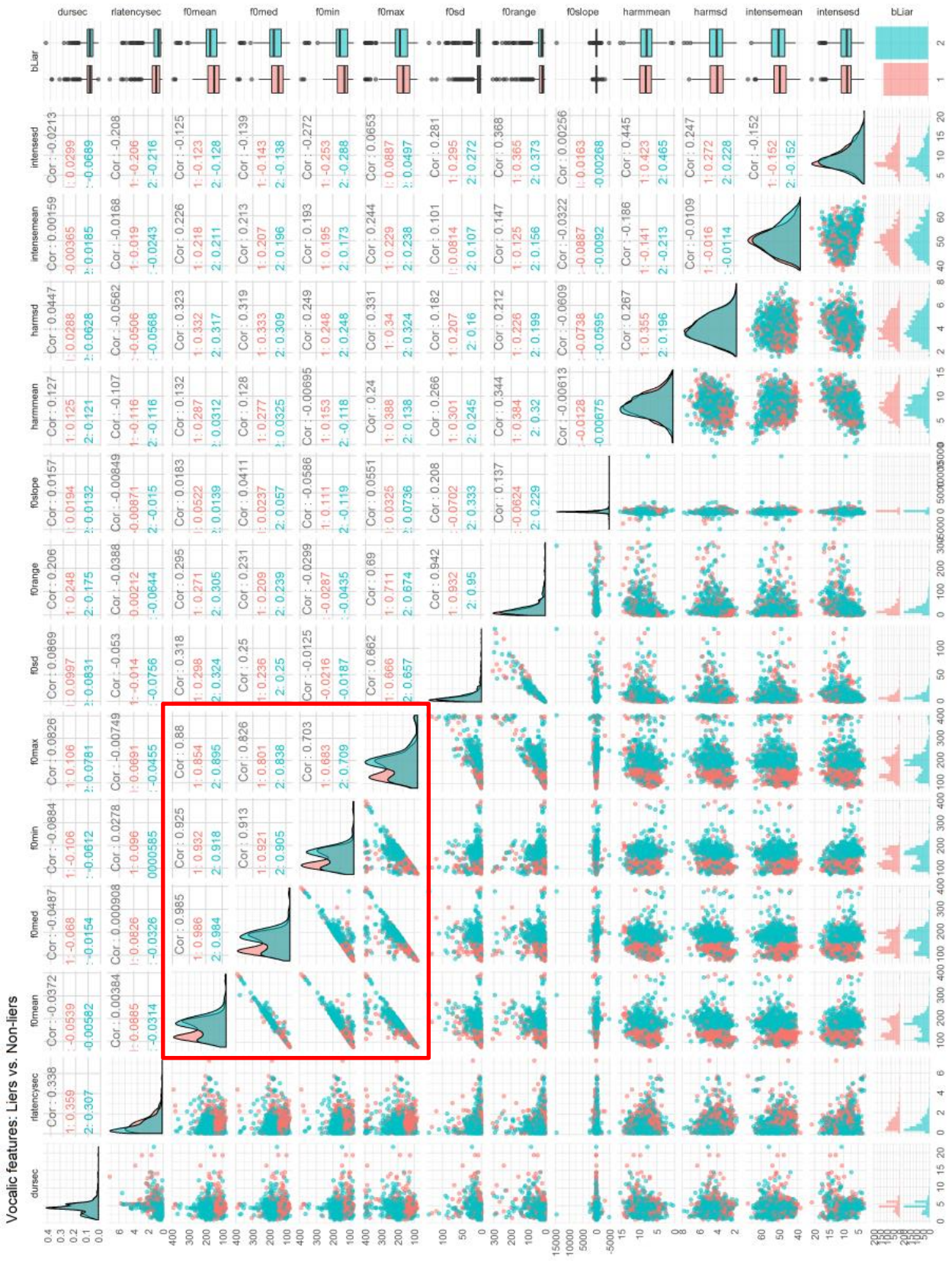


Figure A-1: The values and correlations of vocalic features extracted from third party software, for Liars (bLiar=2) and Non-Liars (bLiar=1). Some features are seen to be affected by being a Liar, some are not. The highly correlated features are also seen and should be removed from further consideration.

Figure 8-2 Scatter plots, histograms and correlations of voice features scores for two opposing classes: Non-Liars (red) vs. Liars (blue), regardless of whether the latter lie or not.

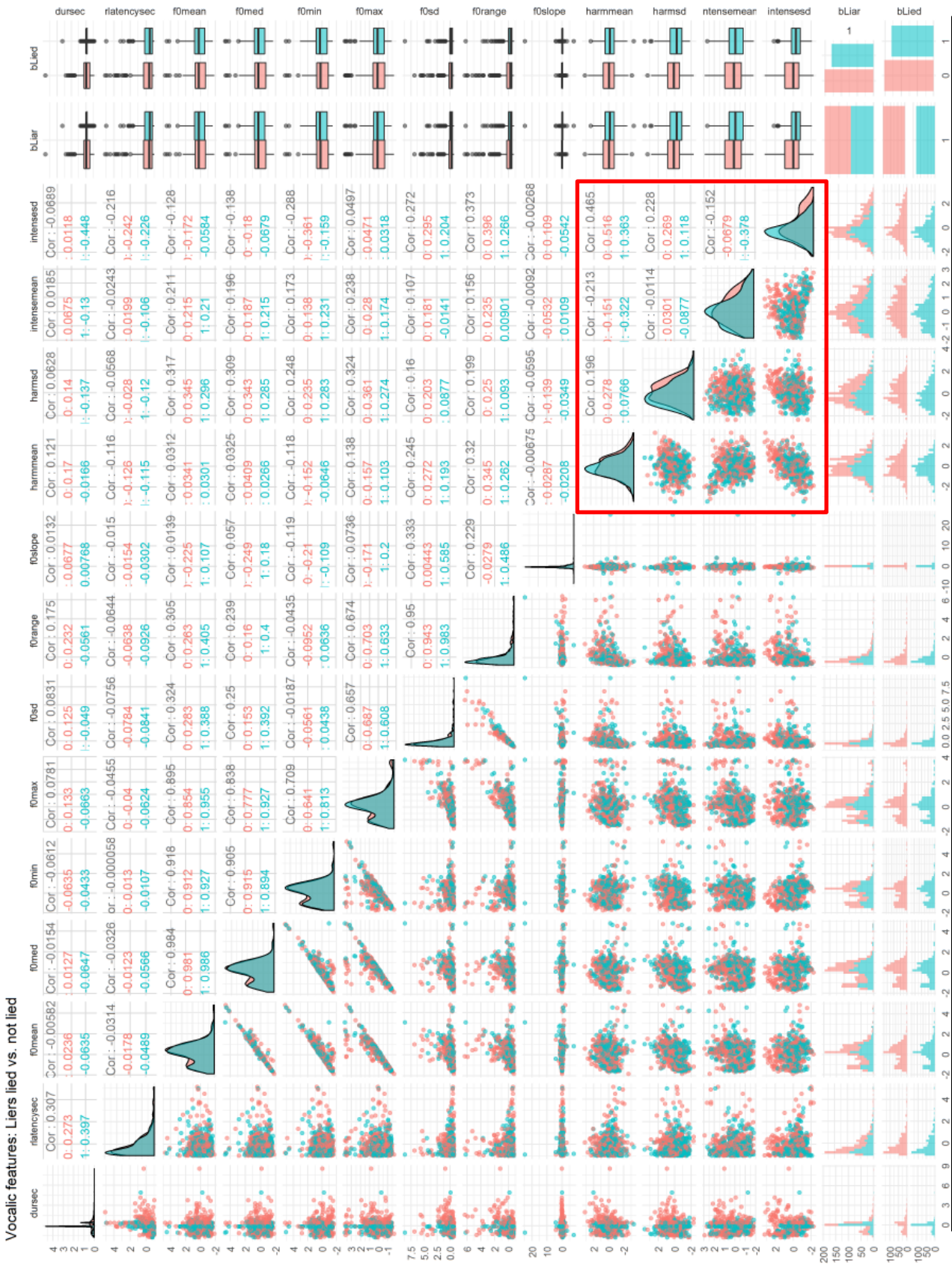


Figure A-2: The *normalized* values and correlations of vocalic features extracted from third party software, for questions where Liars lied (bLied=1, marked Blue) and where they did not lie (bLied=0, marked Red). Features that are affected by lying (found on right side of the graph) are seen to be different from those affected by being a Liar (found on left side of the graph)

Figure 8-3 Scatter plots, histograms and correlations of voice features scores for two opposing classes: Liars not lying (red) vs. Liars lying (blue).

Appendix C. Finding useful features based on two-sample tests and mixed-effects regression

The Task: Find features that are significantly affected by being a Liar and Lying (i.e. by bLiar and bLied variables).
Step 2.3: By using two-sample tests (NB: may not be suitable, if sample size is small and distributions are not normal.
Step 2.4: By using lmer (mixed effect regression). NB: Better models are possible with glmer, nlmer (generalized and nonlinear mixed-effects model regression)

R code

```
analyzeEffectOfLying <- function(dt, cBiometricsUsed, strAffectOf = c("bLiar", "bLied"),
                                 strRandomAffect ="(1 || idAvatar)" ) {

  dtSignificance <- data.table()

  for(i in 1:length(cBiometricsUsed)) {
    biometric <- cBiometricsUsed[i]
    dtRes <- list()
    dtRes$biometric <- biometric

    fml.null <- as.formula(paste0(biometric, "  ~ question + ", strRandomAffect))

    if (strAffectOf=="bLiar") {
      fml.wEffect <- as.formula(paste0(biometric, "  ~ question * bLiar + ",
                                       strRandomAffect))
      model.wEffect <- lmer(fml.wEffect, dt, REML=F)
      model.null <- lmer(fml.null, dt, REML=F)
    } else {
      fml.wEffect <- as.formula(paste0(biometric, "  ~ question * bLied + ",
                                       strRandomAffect))
      model.wEffect <- lmer(fml.wEffect, REML=F, dt[bLiar==1])
      model.null <- lmer(fml.null, REML=F, dt[bLiar==1])
    }
    resAnova<- anova(model.wEffect,model.null);

    dtRes$anovaProb <- 1- resAnova$`Pr(>Chisq)`[2]

    if (strAffectOf=="bLiar") {
      x1 <- unlist(dt[bLiar==1, biometric, with=F])
    }  else {
      x1 <-  unlist(dt[bLiar==1 & bLied==1, biometric, with=F])
    }
    x0 <- unlist(dt[bLiar==0, biometric, with=F])

    dtRes$mean.x1 <- mean(x1, na.rm=T)
    dtRes$mean.x0 <- mean(x0, na.rm=T)
    dtRes$diff.mean <- (dtRes$mean.x1-dtRes$mean.x0) # /(dtRes$mean.x0+dtRes$mean.x1)*2*100

    dtRes$wilcox <- wilcox.test (get(biometric) ~ bLiar, dt, paired=F)$p.value

    dtRes <- dtRes %>% as.data.table()
    dtSignificance <- rbind(dtSignificance, dtRes)
  }

  print(paste0("Using formula `", strRandomAffect,
               "` - Feature [", biometric, "] is effected by [", strAffectOf,
               "] with Probability ",  dtRes$anovaProb
               ". \n It changes its value by ", dtRes$diff.mean
               ". \n In normalized scale: From [", dtRes$mean.x0
               "] to [", dtRes$mean.x1"]")  )
  setkey(dtSignificance,anovaProb); dtSignificance # sort by significance
}
```

Applying the function to select useful features in voice modality

```
strRandomAffect <- "(question || idAvatar)"  # Uncorrelated random intercept and slope

dtSignificance <- analyzeEffectOfLying(dtPupil, cBiometricsUsed,
                                       strAffectOf = "bLiar", # "bLied"
                                       strRandomAffect = "(question || idAvatar)")
print(dtSignificance)
cBiometricsUsed <- dtSignificance[wilcox<=0.025 | anovaProb>=0.95]$biometric
```

The Results of the applying the function with `"bLiar" and "bLied"` are shown and discussed below.

Table 8-1 The best and worst deceit indicators for voice modality computed using mixed effect analysis for the effect of lying (yellow) and being a liar (green).

```
Feature [intensesd] is effected by [bLiar] with Probability 0.1802.
It changes its value by -0.012. In normalized scale: From [0.0065] to [-0.0055]

    biometric anovaProb mean.x1  mean.x0 diff.mean wilcox

 1:   intensesd     0.180 -0.0055  0.00646   -0.0120 0.7026
 2:     f0slope     0.324 -0.0261  0.03068   -0.0568 0.0000
 3: intensemean     0.640  0.0960 -0.11273    0.2087 0.0037
 4:       f0min     0.667  0.1147 -0.13475    0.2495 0.0001
 5:     f0range     0.711  0.0508 -0.05969    0.1105 0.0001
 6:        f0sd     0.747  0.0538 -0.06318    0.1170 0.0000
 7:      f0mean     0.786  0.1450 -0.17033    0.3153 0.0000
 8:    harmmean     0.789 -0.0598  0.07024   -0.1300 0.0590
 9:       f0med     0.798  0.1435 -0.16858    0.3121 0.0000
10:       f0max     0.856  0.1192 -0.14004    0.2593 0.0000
11:      dursec     0.867 -0.0559  0.06569   -0.1216 0.0620
12:      harmsd     0.952  0.0216 -0.02534    0.0469 0.5599
13: rlatencysec     0.968 -0.1148  0.13483   -0.2496 0.0000
```

```
Feature [intensesd] is effected by [bLied] with Probability 1.
It changes its value by -0.247. In normalized scale: From [0.0065] to [-0.2406]"

      biometric anovaProb  mean.x1  mean.x0 diff.mean wilcox

1:       f0med     0.0886  0.17547 -0.16858    0.3441 0.0000
 2:       f0min     0.2048  0.17099 -0.13475    0.3057 0.0001
 3:      f0mean     0.2696  0.16113 -0.17033    0.3315 0.0000
 4:     f0slope     0.3590  0.01237  0.03068   -0.0183 0.0000
 5: rlatencysec     0.6758 -0.18368  0.13483   -0.3185 0.0000
 6:        f0sd     0.7586 -0.01978 -0.06318    0.0434 0.0000
 7:       f0max     0.8935  0.07188 -0.14004    0.2119 0.0000
 8:    harmmean     0.9310 -0.13386  0.07024   -0.2041 0.0590
 9:     f0range     0.9740 -0.07289 -0.05969   -0.0132 0.0001
10:      harmsd     0.9955 -0.12060 -0.02534   -0.0953 0.5599
11: intensemean     0.9998 -0.00873 -0.11273    0.1040 0.0037
12:      dursec     0.9999 -0.18592  0.06569   -0.2516 0.0620
13:   intensesd     1.0000 -0.24056  0.00646   -0.2470 0.7026
```

Result: Mixed-effect analysis of micro-behaviours affected by lying

The success of interview-based credibility assessment, whether done by humans or machines, depends on the knowledge of deceptive signals, which are involuntary micro-behaviours affected by lying. This appendix provides new evidence on the existence of such signals in humans, further classifying them into two types: those affected by lying (i.e., when dishonesty responding to a question) and those affecting by being a liar (i.e., when even honestly responding to a question while hiding some other information). The results are obtained from an automated border control simulation exercise, in which (82) volunteers had to respond to custom declarations questions asked by an automated virtual agent kiosk (AVATAR) equipped with a video camera, eye tracker and a microphone. Biometric features that are measured during a two-minute interview by the kiosk (such as which include proximity, eye-tracking, vocal, facial expression and pupil dynamics information) are analyzed, using mixed-effect regression and two-sample statistical hypothesis tests. The features that provide the best discriminatory power for a deceit detection system are identified in Table 8-1 and Figures below.

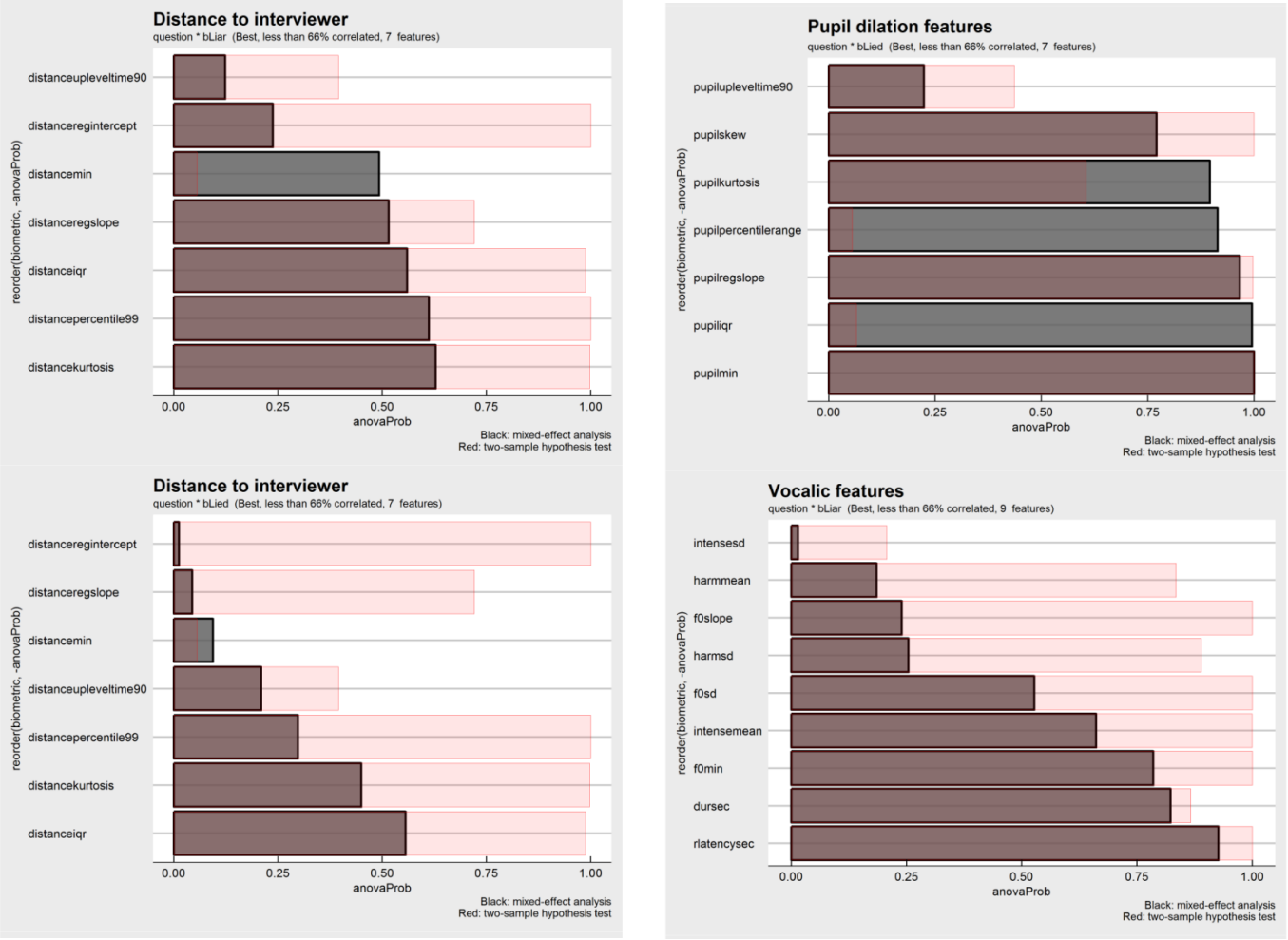


Figure 8-4 The best and worst deceit indicators for proximity modality (left) and pupil modality (right), computed using mixed effect analysis (black) and two-sample hypothesis test (red) for the effect of lying (bottom) and being a liar (top).

Result: An important observation is made - some features are affected by *being a Liar* but not by *Lying*, and vice versa: some features are affected by *Lying*, but not *being a Liar*

The following references were consulted in preparing for this analysis:

- Ieno Zuur et al. Mixed Effects Models and Extensions in Ecology with R. Publisher: Springer New York, 2011.
- Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499. http://arxiv.org/pdf/1308.5499.pdf

- Baayen, R.H.,Davidson,D.J.,Bates,D.M.(2008). Mixed---effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59, 390--- 412.
- Bolker, B.M. , et al (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3), 127---135.
- Pinheiro, J.C. and Bates, D.M. (2000) Mixed-Effects Models in S and SPLUS. Springer
- Molenberghs, G. and Verbeke, G. (2007) Likelihood ratio, score, and Wald tests in a constrained parameter space. Am. Stat. 61, 22–27

Appendix D. Performance of each modality on each interview question

Figure below shows the lie precision accuracy achieved separately on each question by Vocalics , Pupil/Distance and Facial Emotions modality features.

The features were partitioned into 17 question subsets each submitted to a Random Forest classification algorithm. Using 10-fold cross-validation, where each subset is partitioned into 10 random parts, the Random Forest algorithm was tuned to identify the optimal tuning parameters. 10-fold cross-validation creates 10 models holding out one of the partitions as a testing partition or fold. The average Area under the Curve, True Positive and Negative rates of the 10 test fold partitions is reported.

Best and worst performing questions for each modality are seen.

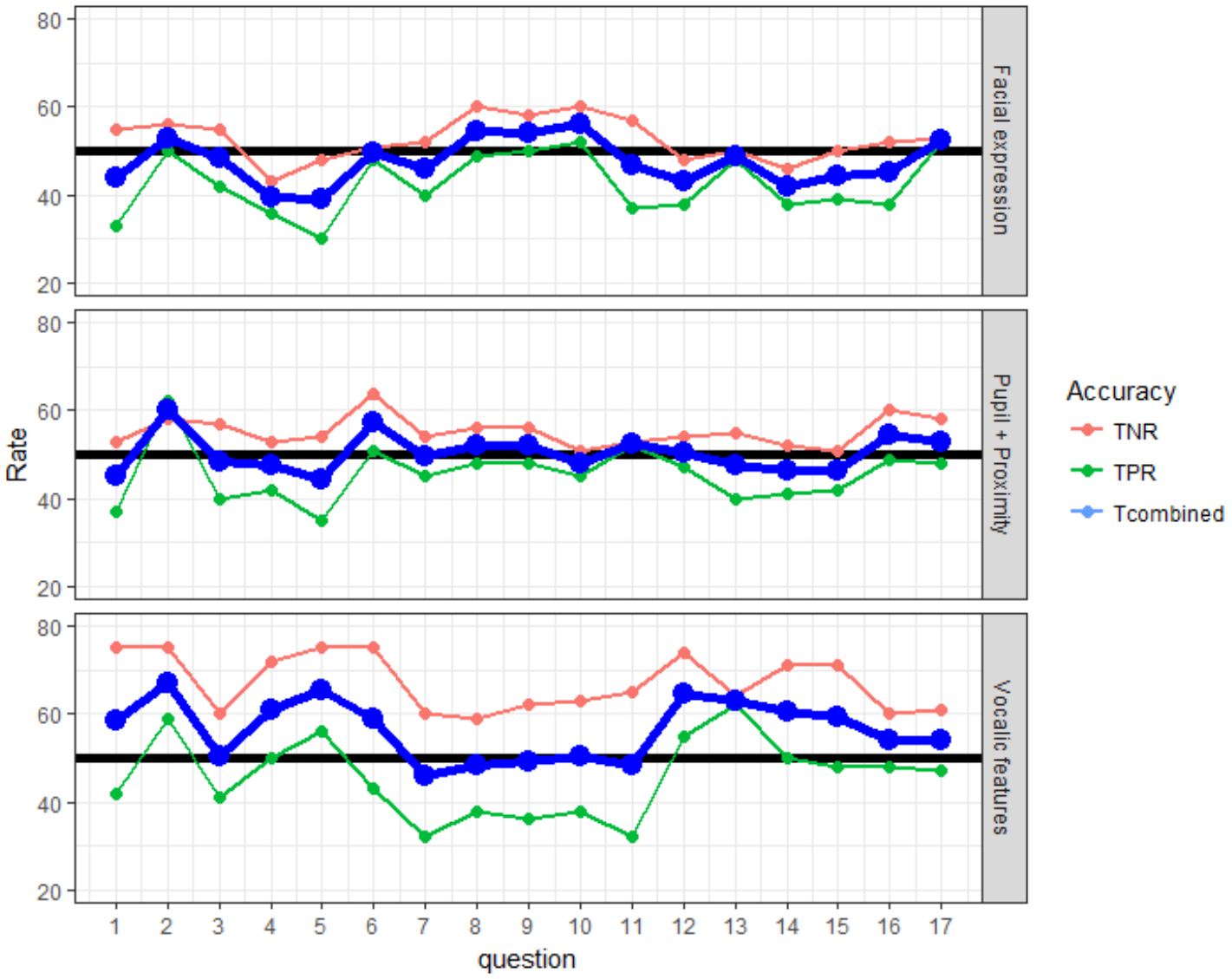


Figure 8-5 Accuracy in detecting lies on each question: by each modality:
From bottom to top each row in an image corresponds to TPR, TNR, and Combined (TPR +TNR)/2.

Appendix E. Interpretable models for each modality

The figures below show the best performing Decision Trees, computed separately for each modality. The TPR is shown in the blue boxes and TNR is shown in green boxes at the bottom of the tree (they need to be added up, when multiple decision branches are used). The numbers appended to the feature names are the question numbers, where the features are measured. Table below summarizes the result.

Table 8-2 Accuracy of interpretable minimalistic decision tree models

	Voice	Pupil & Distance	Eye Fixation Quadrants	Facial Emotions
TNR	56%	51%	67%	46%
TPR	44%	49%	33%	54%

These models are minimalistic, meaning that they used very few features to make decisions. They are however easy to interpret by humans, and their performance, while being worse than that of more complex models, is comparable to that of humans. Using the new insights gained from this report on the best-performing features and questions presented in this report , it is expected that these models can be further improved.

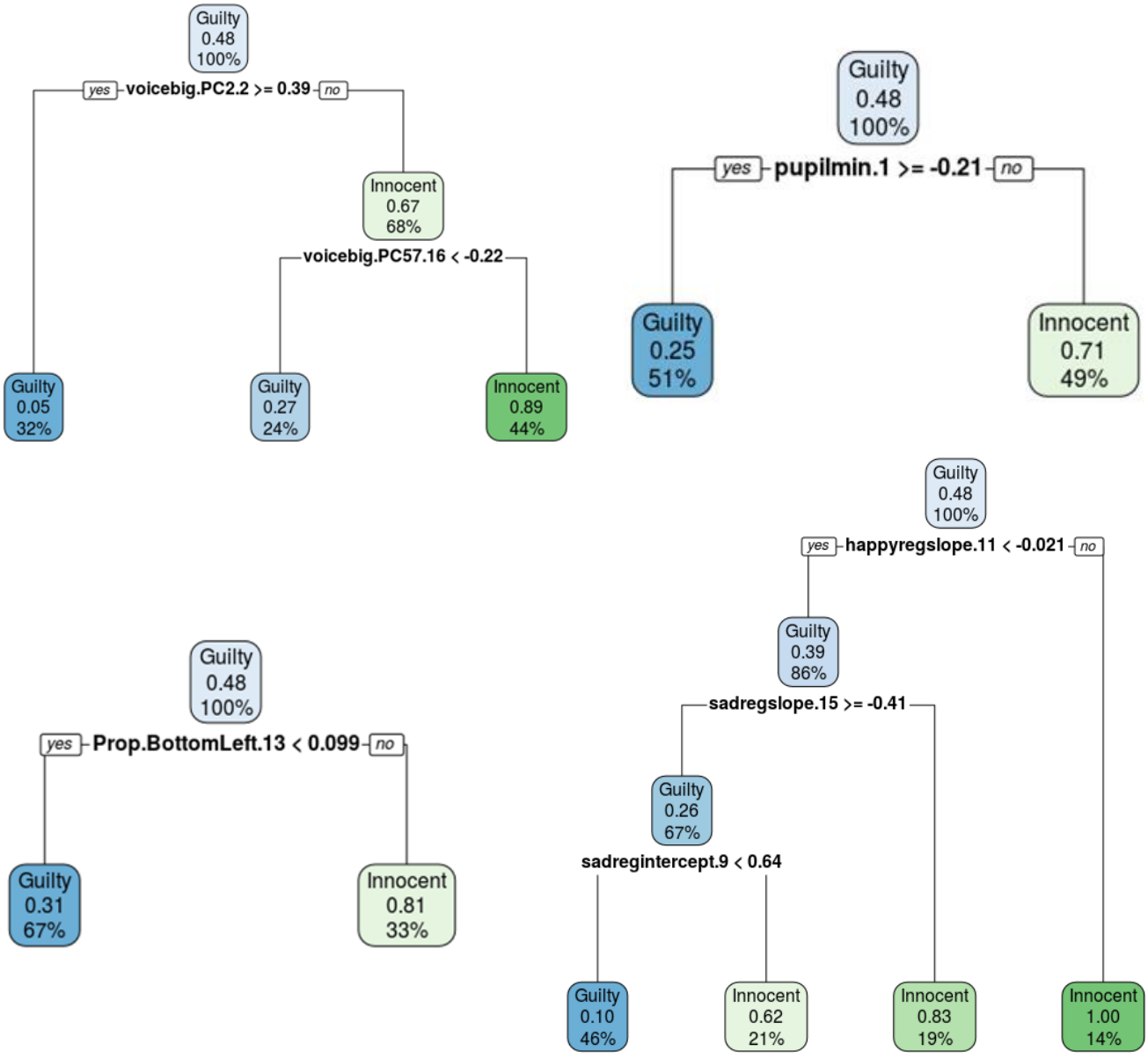


Figure 8-6 Interpretable minimalistic decision tree models built for each modality: from left to right - Vocalics, Pupil/Distance, Eye Fixation Quadrants, Facial Emotions.

Appendix F. Precision-Recall results for each modality

Figure A-4 shows the precision-recall curves computed for four modalities on the dataset collected from the AVATAR experiment at the CBSA.

This analysis was conducted on all of the questions and features in the interview for each modality. Neural network, support vector machine (with non-linear radial kernel), and random forest classification models were parameter tuned and fit using 10-fold cross-validation repeated five times. Repeating the 10-fold cross-validation decreases the variance of the performance measurements and provides a closer approximation of real-world performance. The average Precision-Recall and ROC Area under the Curves were graphed and calculated from the average of the repeated cross-validated test folds.

It is seen that all modalities contribute to lie detection, with the best ones being (in order of performance): voice, Pupil size / Distance to camera, Eye fixation quadrant (which computes the percentage of time the person fixated on each screen quadrant), facial emotions. Precision between 55% and 75% at Accuracy of over 60% is shown to be achievable.

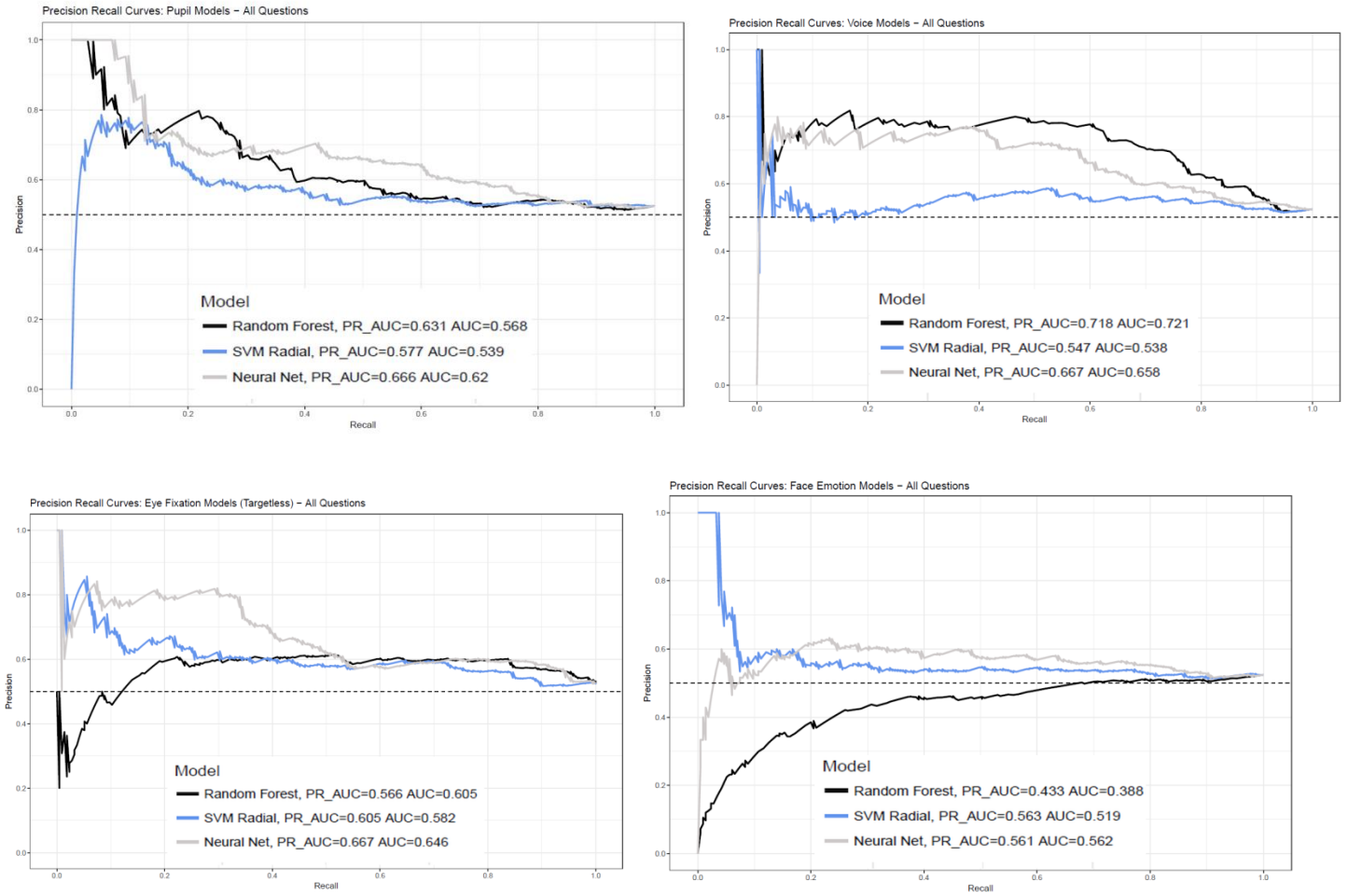


Figure 8-7 Precision vs. Accuracy (Recall) of AVATAR kiosk in detecting Liars by modality: from left to right - Vocalics, Pupil/Distance, Eye Fixation Quadrants, Facial Emotions.