Detecting Concealed Information in Text and Speech

Shengli Hu
Cornell University / Ithaca, New York 14850
sh2264@cornell.edu

Abstract

Motivated by infamous cheating scandals in various industries and political events, we address the problem of detecting concealed information in technical settings. In this work, we explore acoustic-prosodic and linguistic indicators of information concealment by collecting a unique corpus of professionals practicing for oral exams while concealing information. We reveal subtle signs of concealed information in speech and text, compare, and contrast them with those in deception detection literature, thus uncovering the link between concealing information and deception.
We then present a series of experiments that automatically detect concealed information from text and speech. We compare the use of acoustic-prosodic, linguistic, and individual feature sets, using different machine learning models. Finally, we present a multi-task learning framework with acoustic, linguistic, and individual features, that outperforms human performance by over 15%.

1 Introduction

In 2018, a cheating scandal (Mobley, 2018) at the world's most notoriously difficult verbal exam for wine professionals shook the global wine industry — answers were found leaked by some examiners to candidates beforehand — and all results were invalidated; in 2016, with questions leaking ahead of political campaigns (Wemple, 2016), CNN faced a grave scandal from which only more controversies ensued; in 2000, the notorious potential debate leak (Bruni and Van Natta, 2000) in-between the Bush and the Gore campaigns drew the attention of F.B.I. investigators. What all of the three scandals share in common is the fact that it had been difficult to accurately identify who and to whom leaked the critical information, because the party who unfairly obtained the information tried their best to conceal and pretend otherwise.

Despite the importance and potential impact of detecting concealed information, research on detecting concealed information has been scarce. It is partly because large-scale datasets with ground truth labels of information concealment are difficult to come by. It is only in rare cases can we verify the existence of concealed information in the wild.
From the perspective of information attainment and revelation, deception and concealing information are correlated ambiguously. In Table 1, we clarify the difference between the two important concepts with an information grid. When we possess the critical information but appear not in possession, we are concealing information; whereas in contrast, when we do not possess the information but pretend we are in the know, we are deceiving. Despite the proliferation of deception detection studies in text and speech, research on the closely related problem of detecting concealed information has been sparse.

The Information Grid		Appearance	
		Information	No Information
Truth	Information	Honesty	Concealed Information
	No Information	Deception	Honesty

Table 1: The Information Grid: Concealed Information vs. Deception

Prior to trying to detect concealed information, we first ask why might we be able to do so systematically, as opposed to random guessing? Specifically, what makes concealed information detectable? There exist at least two counteracting factors. First, consistent with deception, when individuals are concealing information, they experience potentially greater cognitive load to keep their logic straight, and/or being in fear of being caught, especially when the stakes are high, and

the expectations are great. Second, contrary to deception, because of the endowment with critical information, the candidates also experience more confidence, less fear, and therefore potentially lighter cognitive load, due to the informational advantages. All of these possible offsets make it particularly challenging to control for potential indicators of concealing information.

In the present study, we present a unique corpus of both text and speech, collected from field experiments that provide ground-truth labels, allowing us to initiate the investigation of concealed information, with a focus on technical settings, where some candidates are being evaluated on their technical skills that require logical reasoning. More specifically, we address the following questions:

1. How good (or bad) are humans at detecting concealed information in technical settings?
2. Can we improve on human performance, with a new multimodal dataset, a better understanding of individual differences, and tailored classifiers for audios and texts?
3. How are indicators of concealed information related to those of deception?
4. When are Machine Learning classifiers better (or worse) than human domain experts?

To preview our results, this work contributes to the critical problem of automatic detection of concealed information, increases our scientific understanding of information concealment versus deception and individual differences in concealing information, and presents a series of experiments aimed at automatically detecting concealed information from text and speech. We collect a unique corpus of speech and text from field experiments for the purpose, and show that our multitask learning framework that combines acoustic-prosodic, linguistic, and individual feature sets outperforms baselines by over 11%, and human performance by over 15%.

2 Related Work

There exists limited research in social psychology on Concealed Information Theory (CIT) (Ambach et al., 2010) and interpersonal deception that articulates the nuanced meaning of concealment as a subset of interpersonal deception (Buller et al., 1994). However, ours differ from this body of work in terms of method, scale, and focus. With large-scale computational detection methods based on machine learning and deep learning, we deviate from autonomic and brain electrical measures elicited from small-scale on-campus laboratory experiments and manual analyses.

Besides, as has been detailed in Section 1, the current study is related to deception detection, which has been extensively studied in multiple disciplines such as cognitive psychology, computational linguistics, and paralinguistics, forensic science, etc.

Early work by psychologists (e.g. Ekman et al., 1991, Streeter et al., 1977, Newman et al., 2003) have found indicators of deceptive speech include pitch increases, LIWC (Pennebaker et al., 2001) features, etc. More recently, computer scientists have investigated deception detection in various contexts, identifying cues from texts, speech signals, gestures, and facial expressions. We refer interested readers to Burzo et al. (2017) for an excellent review in this realm.

Language-based indicators of deception have been identified in various contexts. For instance, Bachenko et al. (2008) found that a mixture of linguistic features including hedging, verb tense, and negative expressions are predictive of truthfulness in criminal narratives, interrogation, and legal testimony. Ott et al. (2011) investigated online deceptive opinion spams by crowdsourcing a dataset of fake hotel reviews using Amazon Mechanical Turk, and found deceptive spams exhibit more positive emotions, first-person singulars, concrete expressions, and fewer spatial configurations. Studies in the similar vein include Hancock et al. (2007), Mihalcea and Strapparava (2009), Feng et al. (2012), etc. Toma and Hancock (2010), Guadagno et al. (2012), Joinson and Dietz-Uhler (2002) and Warkentin et al. (2010) explored deception detection in more diverse online settings such as online dating, social networks, and online communities.

There has also been much progress in identifying cues of deception in speech signals. Levitan et al. (2015) collected a large-scale corpus of cross-cultural speech of deception and truth-telling, coupled with individual features such as personality traits. They found that gender, native language, and personality information significantly improves classification accuracy along with acoustic-prosodic features. Levitan et al. (2016)

combined acoustic-prosodic, lexical, and phonotactic features to automate deception detection and outperformed human performance by a large margin. Levitan et al. (2018a) and Levitan et al. (2018b) tested for statistically significant acoustic-prosodic and linguistic indicators of deception detection. Moreover, Mendels et al. (2017) trained a hybrid deep learning model that combines speech signals with textual features, outperforming shallow machine learning methods.

Videos of deceptive and non-deceptive speech have also been collected to leverage the visual cues for automatic detection. Pérez-Rosas et al. (2015a), and Pérez-Rosas et al. (2015b) collected real-life trial videos and applied image processing methods to extract gestures and facial expressions, which prove to improve the performance of deception detection classifiers.

Ideologically, the current study is related to studies that explore how to improve human decision-making processes with machine learning algorithms, especially in tasks that prove difficult for humans. Other research efforts include Kleinberg et al. (2017) where algorithms aid judges' decisions, Ranganath et al. (2009) where machines detect flirtation better than humans, and the deception detection literature reviewed above where machines outperform humans by a large margin.

3 Data

3.1 Blind Tasting Game

We design a field experiment that mimics the setting of the motivating cheating scandal in blind tasting oral exams (in Section 1) and provides financial incentives to participants to conceal critical information, which is randomly assigned to individuals.

More specifically, in each session of the blind tasting game, there are $5-10$ wine professionals participating in $7-15$ rounds sequentially. During each round, one mysterious wine, the identity of which is known to one participant by chance, is poured. Every individual including the informed one, proceeds to taste, describe, reason, and conclude on his/her guesses about the identity of the wine, in a random sequence, both verbally and in writing. The professionals participate to practice their tasting skills, and strive to make as many correct calls about wines' identities as possible. Once every individual has voiced their opinions, the identity is revealed, and each participant is asked to provide guesses of the informed participants before revelation — they can write as many as they wish. The participants who have done the best job concealing information (i.e. the least correct guesses of concealing information by others) in aggregate are rewarded with fine wines as incentives at the end of each session.

We recorded 49 sessions with a total of 41 professionals in rooms with soundproof equipment and collected answer sheets on which participants wrote descriptions, conclusions, and guesses of informed individuals.

We also collected information about participants' gender, native language, credentials, and granular domain knowledge (self-confidence in identifying every style or region of wines) in post-session questionnaires. 88% of participants' native language is American English, the rest include Korean, Spanish, British English, and Chinese. 61% of participants have passed the level of certified sommelier or above with the Court of Master Sommelier, one of the most authoritative institutions in the industry, especially in the United States; and 41% of participants have passed the third level with the Wine $\&$ Spirit Education Trust, the other most authoritative institution in Europe.

3.2 Pre-processing

We manually annotated the audio samples by speaker, with or without information, and the identity of wine, using Praat (Boersma et al., 2002). We discarded audios unrelated to the tasting game, such as small talk. The resulting audio samples were then transcribed with wit.ai API for automatic speech recognition, and hand-corrected afterward. We also transcribed the written answers and annotated accordingly.

The speech was tagged and aligned with the speaker id. We then segmented it into turn units, where a turn is defined as a maximal sequence of inter-pausal units (pause-free segments separated by a minimum pause length of 50 ms) from a single speaker without any interlocutor speech that is not a backchannel. Labels of speaker id, wine identity, speakers' guesses of both wine and informed individuals were assigned to each turn accordingly. We define single turn segments as individual turns of a speaker in any round separately and aggregate them by speaker and round as multiple turn segments. Our classification is performed on both segmentations of the data, whereas statis-

tical analyses are on multiple turn segments.
The resulting corpus totaled 164 hours, and 3288 multiple turn and 9104 single turn segments. We randomly split our entire set into training, development, and testing sets at the ratio of 70:10:20 separately for single and multiple turn. Evaluation results were based on 5-fold cross-validation.

4 Feature Extraction

4.1 Acoustic-prosodic Features and Indicators

We extract 8 low-level acoustic features commonly studied in speech research: intensity mean and max, pitch mean and max, 3 voice quality features (shimmer, jitter, noise-to-harmonics ratio), and speaking quality, as well as 13 Mel-Frequency Cepstral Coefficients (MFCCs) per window of 256 frames and stride of 100 frames, using Praat (Boersma et al., 2002), Parselmouth (Jadoul et al., 2018), and python speech features library.
Following previous studies on deception, we use openSMILE (Eyben et al., 2010) to extract two feature sets from the InterSpeech challenges: the 2013 Computational Paralinguistics Challenge baseline feature set (IS2013) (Schuller et al., 2013), and the 2009 Emotion challenge baseline feature set (IS2009) (Schuller et al., 2009). The two feature sets contain 6373 and 384 features respectively, from computation of various functionals over low-level descriptors such as pitch (fundamental frequency), intensity, spectral, cepstral, duration, voice quality, spectral harmonicity, and psychoacoustic spectral sharpness. These have been shown useful for many tasks such as native language detection (e.g., Keren et al., 2016), emotion detection (e.g., Eyben et al., 2013, Satt et al., 2017), sincerity (e.g., Zhang et al., 2016, Herms, 2016), and deception detection (e.g., Zhang et al., 2016, Herms, 2016). The two feature sets were used in our machine learning classification tasks. All the audio features are z-score normalized by speaker.

Table 2 shows the statistically significant low-level acoustic features (marked by S) for both classes based on paired t-tests between the features of truthfulness and information concealment, corrected for family-wise Type I error by controlling the false discovery rate (FDR) at $\alpha = 0.05\%$. (S) indicates significant uncorrected p-values.
As is shown in Table 2, across all speakers (the last column), we observe an increase in maximum pitch, intensity, speaking rate, and a decrease in duration, suggesting that speakers on average tend to speak with higher maximum pitch, intensity, rate, and shorter duration when concealing information. It has been documented in multiple deception detection studies (e.g. Levitan et al., 2018a) that people also tend to speak with a higher maximum pitch and intensity when telling a lie.
To understand the individual differences in speech with concealed information, we report the same test statistics for specific subsets of speakers — grouped by gender, and skill level. We find that maximum pitch is significantly increased in information concealment for male speakers but not for female speakers, and that increased speaking rate in information concealment for speakers with lower skill. These results largely echo the results in recent deception detection studies in interview dialogues (e.g. Levitan et al., 2018a), except that we found the total duration was longer for truthful speech than speech with concealed information. The finding about increased speaking rate in relatively lower-skilled professionals supports the hypothesis that extra information boosts confidence level and outweighs the effect of increased cognitive load when concealing information.

Feature	Male	Female	Low Skill	High Skill	All
Pitch (max)	S				S
Pitch (mean)					
Intensity (max)	S	S	(S)		S
Intensity (mean)		(S)			
Speaking Rate			S		S
Duration		(-)(S)	(-)(S)		(-)(S)
Voice Quality					

Table 2: Low-level Acoustic Indicators of Information Concealment

4.2 Linguistic Features and Indicators

LIWC: previous research in speech and text found LIWC dimensions useful for predicting personality (Newman et al., 2003), deception (Levitan et al., 2018b), etc., therefore we extract 93 semantic classes using LIWC 2015 (Pennebaker et al., 2001, Pennebaker et al., 2015). They include standard linguistic dimensions (e.g., *pronoun*, *article*), grammar (e.g., *verb*, *adj*, *compare*), psychological processes (e.g., cognitive process *cogproc*, social processes *social*, affective processes *affect*), time orientation (e.g., *focuspast*, *focuspresent*, *focusfuture*), relativity (*relativ*),

and formality (e.g., informal language *informal*, *Netspeak*).

Linguistic: we extract 10 linguistic features based on results from previous literature. Included are binary and numeric features capturing hedging (Choi et al., 2012, Prokofieva and Hirschberg, 2014), linguistic and syntactical distinctiveness, subjectivity, sentiment (valence, intensity) (Pang et al., 2008), contraction, level of detail (Li and Nenkova, 2015), and contextual concreteness.

We measure hedging using a rule-based algorithm introduced in Prokofieva and Hirschberg (2014), complemented with a comprehensive hedging dictionary released by Choi et al. (2012).

We measure linguistic and syntactical distinctiveness by training a "common language model" using a wine review corpus that consists of $860,119$ reviews from four major websites — Vinous, Wine Spectator, Wine Enthusiast, and Decanter, as summarized in Hu (2018). Following Danescu-Niculescu-Mizil et al. (2012), we use unigrams, bigrams, and trigrams for training and use the model to predict the likelihood of given sentences in our corpus, which measures linguistic distinctiveness. For syntactical distinctiveness we add Part-of-speech tags to the language model.

We measure contextual concreteness by building a domain-specific rule-based algorithm following the tasting grid widely used by wine professionals. Our method counts both the number and percentage of cluster descriptor versus item descriptor, weighted by pre-specified weights calculated based on a weighting scheme identical to tf-idf except that the document corresponding to representative descriptors of a style and region of wine.

Subjectivity and sentiment measures were extracted with TextBlob (Loria, 2010).

Length: we include the average number of words by turn and sentence, the average length of words by turn, sentence, and word.

Ngrams: we extract unigrams, bigrams, and trigrams, which has been shown useful for domain-specific deception detection (Ott et al., 2011).

Embeddings: we obtain distributed representations of words to capture semantic relationships using GloVe (Pennington et al., 2014) word vectors trained on 1B tweets and the same wine review corpus used for the common language model.

All the linguistic features are extracted for both audio transcriptions and written texts. We calculate the differences in-between using Euclidean distance (where Ngrams are preprocessed to be binary on every dimension).

Table 3 shows (1) the top ngram features for both concealed information and truthful classes from a logistic regression classifier, which yields an F1-score of 60.13% with minimal manipulation; (2) the statistically significant LIWC, linguistic, and other features for both classes based on the same tests as detailed in Section 4.1, by taking the union of significant feature sets from transcriptions and written texts. We further compare our results with those in recent deception detection studies (Levitan et al., 2016, Levitan et al., 2018b, Levitan et al., 2018a). In Table 3, we denote significant features consistent with deception literature as red and underlined, and those opposite with deception literature as *blue and italicized*.

Feature	Concealed Information	Truthful
N-grams	*yeah*, but it, citrus, correct, ruby, did not, lift, botrytis, would not	*uh um*, there is, there are, was like, so, slight, not sure, blossom, clear
LIWC	*clout*, certain, function, cogproc, negate, *discrep*, differ, assent, posemo	compare, pronoun, verb, *ingest*, feel
Syntax	*adj*, *adverb*, syn_distinct	
Else	*specificity*, $\Delta(Trans, Text)$	hedging, #word, length

Table 3: Linguistic Indicators of Information Concealment vs. Truthfulness

Consistent with Benus et al. (2006), we found that the use of filler pauses such as "um" were correlated with truthful speech. The LIWC *cogproc* (cognitive processes — e.g. "cause", "think", "know"), *certain*, *posemo* (positive emotion), *negation*, and *assent* features were significantly more frequent in speech with concealed information, in line with Levitan et al. (2018b), supporting the hypotheses that cognitive load, as well as confidence level, increases with the pressure of concealing information.

We also found the LIWC *compare*, *verb*, and *feel* features, backed by hedging, total word count and length significantly more frequently associated with speech without concealed information, suggesting an interesting balance of more visceral responses and deliberation associated with truth-telling in technical settings.

Other significant indicators of concealed in-

formation include syntactical distinctiveness (syn_distinct), specificity, clout, discrepancy, and disparity between speech and written text ($\Delta(Trans, Text)$), the results regarding clout (confidence) and discrepancy are consistent with Levitan et al. (2018b).

Some of the ngrams features appear to echo the other linguistic indicators. For instance, "botrytis" is a precise winemaking term that is usually associated with aromas of honey, ginger, and saffron, which corresponds to the specificity feature identified as an indicator of information concealment; "clear" is a fairly general term in wine talk that indicates neither certainty nor specificity, and therefore more indicative of truthfulness; the appearance of the word "ruby" is in accordance with the statistics that wine professionals in our sample performed better on and are more confident in calling red wines than white wines.

5 Classification Experiments

We first balance our dataset by random upsampling, since the number of negative labels is more than twice of positive labels.

5.1 Baseline Models

We trained Logistic Regression (LR) classifiers with ngrams, and Random Forest (RF) classifiers with acoustic features as baseline models for text and speech respectively. For LR, we varied preprocessing methods (stop words, number of ngrams, binary vs. numeric), and the most performant LR model uses only bigram features. For RF, we varied the number of trees, the choice of feature sets detailed in Section 4.1. The most performant RF model uses 800 tree estimators, and the IS 2009 Emotion Challenge feature set alone.

5.2 Deep Learning Models

Given the results from baseline models and previous literature (Mendels et al., 2017; Levitan et al., 2016), we train Bidirectional Long Short-Term models (BiLSTM, Schuster and Paliwal, 1997; Zhang et al., 2015) with sequences of word embeddings, Multi-Layer Perceptrons (MLP) with acoustic feature sets, and the combinations thereof. The GloVe embeddings were used to initialize the weights but back propagation was also allowed to update embedding values during training. We use Bayesian optimization (Snoek et al., 2012) to tune the hyperparameters. It was used to maximize the F1 scores on the development set, based on various hyperparameters including learning rate, number of hidden layers of MLP, the number of hidden units per layer, optimizers and associated parameters, dropout rate, and batch size. and concatenate embeddings learned from acoustic features passed through an MLP and those passed through a BiLSTM for the last softmax layer. The combined model structure follows Mendels et al. (2017) except that we used 4 hidden layers for MLP, and concatenated additional individual features before the last softmax layer. Our model consists of four fully connected layers, each with 680 hidden units followed by ReLU (Krizhevsky et al., 2012) activations. We use a softmax layer with two outputs that corresponds to the two classes (Concealment vs. Truthful) in our task, trained on categorical cross-entropy as the loss function. Training process also includes Batch Normalization (Ioffe and Szegedy, 2015) and Dropout (Srivastava et al., 2014) implementation (keep probability being 0.6) upon the output of each layer. The optimizer is Adam (Kingma and Ba, 2014). The BiLSTM trained on word embeddings is then merged with MLP and individual feature vectors by concatenation based on last hidden layers. The base model with trigrams did perform slightly better than BiLSTM with GloVe with an improvement of 0.50 and 0.31 of F1 scores over bigrams but the resulting vector dimension does not balance well with that from MLP and individual features for gradient propagation. Therefore, we use BiLSTM with GloVe of comparable performance and greater simplicity. To prevent the acoustic MLP from being penalized more than the linguistic BiLSTM, we adopt an auxiliary softmax layer to the BiLSTM output concatenated with individual features, with a parameter chosen to be 0.41 by Bayesian optimization.

5.3 Multi-task Learning

Based on the combined model in Section 5.2, we explore multi-task learning by adding two more tasks that share the same training set, and an additional dataset scraped from blind tasting video and audio clips posted by Guild of Sommeliers. It consists of 5.5 hours clean blind tasting demonstrations, and 21 rounds. The two additional tasks are predicting if the speaker's

answer is correct, and the identity of the wine. The overall structure is shown in Figure 1.

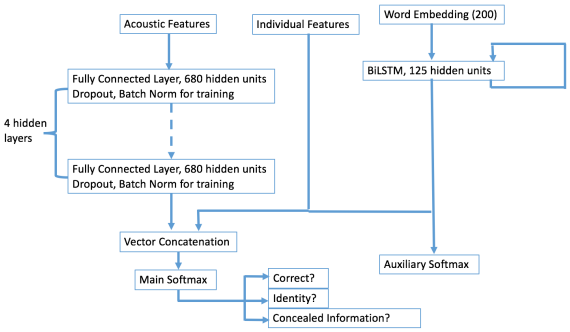

Figure 1: Multi-task Leaning Framework Combining Acoustic, Linguistic, and Individual Features

5.4 Measuring Human Performance

We obtained the overall human performance at an F1-score of 56.28 for multiple turn segmentation, by converting individual guesses into binary labels and aggregating them. There is no statistics for single turn segmentation available because we asked participants to provide guesses at the end of each round. The F1-score of individuals with higher credientials did beat Random Forest with IS 2009 features at 63.14. We speculate that it could be because these participants more versed in blind tasting are also better at detecting inconsistencies between descriptions and conclusions, which are telltale signs of concealed information in our context.

If we were to obtain human performance for single turn segmentation, we would have to ask the participants to provide guesses every time there was a speaker turn, which would greatly disrupt the experiment. While it is entirely possible to hire another group of qualified wine professionals familiar with the setting to help annotate the dataset by both single and multiple turn, and we could have obtained human performance by asking other wine professionals not involved in the experiments to listen to the audio clips and provide labels, but it was not feasible due to logistic and financial constraints at the time of implementation.

Since some of the professionals know one another well and the task of guessing several individuals out of all for each round makes it easier than the detection task faced by algorithms, due to social and order effects (the sequence of rounds was randomly determined to counterbalance order effects from treatments, but it does not eliminate the biases from human performance measures), we argue that our statistics for human performance is biased, but it provides the upper bound because the real performance could be even worse.

5.5 Results

Table 4 shows the F1-scores of the most performant models from each model class as described in Section 5. We also marked human performance on multiple turn segments in red in the first row, since it is worse than all the models in the table. Across all the models and feature sets, multiple turn segmentation yields better F1-scores compared to single turn segmentation. It is intuitive in the sense that, multiple turn segments contains more information than single turn segments, leading to more informative features that help classification. Consistent with Levitan et al. (2015), we have also found individual features such as gender, skill level, and native language, boost classification performance by a relatively large margin — the same magnitude as the boost from combining acoustic and linguistic features. Furthermore, multi-task learning with auxiliary classifiers boosts F1-score by the same margin as adding individual features to the joint model, which is higher than human domain experts' performance by 15.23%.

Model	Features	F1 (single / multiple turn)
Logistic Regression	Bigrams	Human: NA / 56.28 61.18 / 65.45
Random Forest	IS 2009	59.23 / 60.03
MLP	IS 2009	63.96 / 67.27
BiLSTM	GloVe	61.41 / 67.35
MLP + BiLSTM	IS 2009, GloVe	64.12 / 68.57
MLP + BiLSTM	IS 2009, Individual Features, GloVe.	64.14 / 70.02
MLP + BiLSTM + Multi-task	IS 2009, Individual Features, GloVe.	65.16 / 71.51

Table 4: Classification Results of Baselines, DL Models, Combined DL Model, and Multi-task Learning Model

6 Conclusions, Limitations, and Future Directions

We have presented a study of concealed information in text and speech. Our analysis of acoustic-prosodic and linguistic characteristics of informa-

tion concealment, contrasted with those of deception, provides insight into the nature of text and speech with or without concealed information. We have also evaluated the performance of several machine learning classification methods to the critical problem of detecting concealed information in technical settings. We developed a hybrid multi-task learning model that outperforms baseline models by 11.48% and human domain experts by 15.23%.

The current study is by no means perfect. First, more samples and machine learning experiments could have been done, had we had more time, funding, and resources. Second, during the blind tasting games, the identity of each wine during each round was known to one participant by chance, in that, every potential grape-region combo in a total pool of 38 combos was randomly assigned to one participant beforehand without replacement. At the time of pouring, there was no secret mechanism in place to inform the particular participant who brought the wine being poured. However, given the fact that (1) it was common knowledge (each one knows it, if each one knows that the others know it, if each one knows that each one knows that the others know it, and so on) that each grape-region combo was assigned to no one or one individual for each session, and (2) all participants were required to bring a wine of the assigned grape-region that they know very well and ensure it of a classical style most representative of the grape and region, the task of detecting self-brought wines becomes trivial to our participants, and therefore the informing mechanism stands. We acknowledge that a cleaner and cleverer design could have been implemented to randomly assign and secretly inform participants, however, that would require hiring more independent administrators and sacrificing some participants' practice opportunities, which were the reasons why we settled for the current setting. Third, more analyses of acoustics such as pitch and tonal contour, phonotactic variations could be incorporated to further explore the space of information concealment in speech. Fourth, additional experiments with identified significant acoustic and linguistic features would add more weight to the current paper.

Lastly, we look forward to further exploring this line of research by investigating:

1. the individual differences in both detecting concealed information and concealing information, by analyzing the features across groups defined by individual personality traits (Fornaciari et al., 2013, An et al., 2018), ethnics, native languages, and different dimensions of professional skills;
2. the result and model robustness by collecting and testing other field data such as board games;
3. the predictive power of phonotactic variation features;
4. the relationship between perceived information concealment and concealing information;
5. how *soon* can we detect concealed information;
6. how to conduct domain adaptation with regards to detecting concealed information;
7. efficient ways to make the multi-task learning framework scalable.

In addition, this line of work might inspire new methods for detecting insider trading in financial markets.

References

Wolfgang Ambach, Stephanie Bursch, Rudolf Stark, and Dieter Vaitl. 2010. A concealed information test with multimodal measurement. *International Journal of Psychophysiology* 75(3):258–267.

Guozhen An, Sarah Ita Levitan, Julia Hirschberg, and Rivka Levitan. 2018. Deep personality recognition for deception detection. *Proc. Interspeech 2018* pages 421–425.

Joan Bachenko, Eileen Fitzpatrick, and Michael Schonwetter. 2008. Verification and implementation of language-based deception indicators in civil and criminal narratives. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, pages 41–48.

Stefan Benus, Frank Enos, Julia Bell Hirschberg, and Elizabeth Shriberg. 2006. Pauses in deceptive speech. *academiccommons.columbia.edu* .

Paul Boersma et al. 2002. Praat, a system for doing phonetics by computer. *Glot international* 5.

Frank Bruni and Don Van Natta. 2000. The 2000 campaign: The inquity; f.b.i. widens investigation into debate leak. *The New York Times* .

David B Buller, Judee K Burgoon, Cindy H White, and Amy S Ebesu. 1994. Interpersonal deception vii: Behavioral profiles of falsification, equivocation, and concealment. *Journal of language and social psychology* 13(4):366–395.

Mihai Burzo, Mohamed Abouelenien, Veronica Perez-Rosas, and Rada Mihalcea. 2017. Multimodal deception detection. *The Handbook of Multimodal-Multisensor Interfaces* 2.

Eunsol Choi, Chenhao Tan, Lillian Lee, Cristian Danescu-Niculescu-Mizil, and Jennifer Spindel. 2012. Hedge detection as a lens on framing in the gmo debates: A position paper. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*. Association for Computational Linguistics, pages 70–79.

Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. 2012. You had me at hello: How phrasing affects memorability. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, pages 892–901.

Paul Ekman, Maureen O'Sullivan, Wallace V Friesen, and Klaus R Scherer. 1991. Invited article: Face, voice, and body in detecting deceit. *Journal of nonverbal behavior* 15(2):125–135.

Florian Eyben, Felix Weninger, and Björn Schuller. 2013. Affect recognition in real-life acoustic conditions-a new perspective on feature selection. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, pages 1459–1462.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, pages 171–175.

Tommaso Fornaciari, Fabio Celli, and Massimo Poesio. 2013. The effect of personality type on deceptive communication style. In *2013 European Intelligence and Security Informatics Conference*. IEEE, pages 1–6.

Rosanna E Guadagno, Bradley M Okdie, and Sara A Kruse. 2012. Dating deception: Gender, online dating, and exaggerated self-presentation. *Computers in Human Behavior* 28(2):642–647.

Jeffrey T Hancock, Lauren E Curry, Saurabh Goorha, and Michael Woodworth. 2007. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes* 45(1):1–23.

Robert Herms. 2016. Prediction of deception and sincerity from speech using automatic phone recognition-based features. In *Interspeech*. pages 2036–2040.

Shengli Hu. 2018. Somm: Into the model. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. pages 1153–1159.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* .

Yannick Jadoul, Bill Thompson, and Bart De Boer. 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics* 71:1–15.

Adam N Joinson and Beth Dietz-Uhler. 2002. Explanations for the perpetration of and reactions to deception in a virtual community. *Social Science Computer Review* 20(3):275–289.

Gil Keren, Jun Deng, Jouni Pohjalainen, and Björn W Schuller. 2016. Convolutional neural networks with data augmentation for classifying speakers' native language. In *INTERSPEECH*. pages 2393–2397.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human decisions and machine predictions. *The Quarterly Journal of Economics* 133(1):237–293.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. pages 1097–1105.

Sarah Ita Levitan, Guozhen An, Min Ma, Rivka Levitan, Andrew Rosenberg, and Julia Hirschberg. 2016. Combining acoustic-prosodic, lexical, and phonotactic features for automatic deception detection. In *INTERSPEECH*. pages 2006–2010.

Sarah Ita Levitan, Michelle Levine, Julia Hirschberg, Nishmar Cestero, Guozhen An, and Andrew Rosenberg. 2015. Individual differences in deception and deception detection. *Proceedings of Cognitive* .

Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg. 2018a. Acoustic-prosodic indicators of deception and trust in interview dialogues. *Proc. Interspeech 2018* pages 416–420.

Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg. 2018b. Linguistic cues to deception and perceived deception in interview dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). volume 1, pages 1941–1950.

Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

S Loria. 2010. Textblob, python library for processing textual data., 2017. *URL: https://web. archive. org/web/20161222050214/http://textblob. readthedocs. io/en/dev* 2.

Gideon Mendels, Sarah Ita Levitan, Kai-Zhan Lee, and Julia Hirschberg. 2017. Hybrid acoustic-lexical deep learning approach for deception detection. In *INTERSPEECH*. pages 1472–1476.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Association for Computational Linguistics, pages 309–312.

Esther Mobley. 2018. Why a cheating scandal is shaking the sommelier world. *San Francisco Chronicle* .

Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin* 29(5):665–675.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 309–319.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2(1–2):1–135.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report, what is this.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71(2001):2001.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. pages 1532–1543.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015a. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, pages 59–66.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, Yao Xiao, CJ Linton, and Mihai Burzo. 2015b. Verbal and nonverbal clues for real-life deception detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pages 2336–2346.

Anna Prokofieva and Julia Hirschberg. 2014. Hedging and speaker commitment. In *Proceedings of the 5th International Workshop on Emotion, Social Signals, Sentiment & Linked Open Data, Reykjavik, Iceland*. pages 10–13.

Rajesh Ranganath, Dan Jurafsky, and Dan McFarland. 2009. It's not you, it's me: detecting flirting and its misperception in speed-dates. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, pages 334–342.

Aharon Satt, Shai Rozenberg, and Ron Hoory. 2017. Efficient emotion recognition from speech using deep learning on spectrograms. In *INTERSPEECH*. pages 1089–1093.

Björn Schuller, Stefan Steidl, and Anton Batliner. 2009. The interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*.

Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*. pages 2951–2959.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15(1):1929–1958.

Lynn A Streeter, Robert M Krauss, Valerie Geller, Christopher Olson, and William Apple. 1977. Pitch changes during attempted deception. *Journal of personality and social psychology* 35(5):345.

Catalina L Toma and Jeffrey T Hancock. 2010. Looks and lies: The role of physical attractiveness in online dating self-presentation and deception. *Communication Research* 37(3):335–351.

Darcy Warkentin, Michael Woodworth, Jeffrey T Hancock, and Nicole Cormier. 2010. Warrants and deception in computer mediated communication. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, pages 9–12.

Erik Wemple. 2016. With question-leaking, cnn has a scandal on its hands. *The Washington Post* .

Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. 2015. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*. pages 73–78.

Yue Zhang, Felix Weninger, Zhao Ren, and Björn W Schuller. 2016. Sincerity and deception in speech: Two sides of the same coin? a transfer-and multi-task learning perspective. In *INTERSPEECH*. pages 2041–2045.