JOINT LEARNING OF CONVERSATIONAL TEMPORAL DYNAMICS AND ACOUSTIC FEATURES FOR SPEECH DECEPTION DETECTION IN DIALOG GAMES

Huang-Cheng Chou[1,2], *Yi-Wen Liu*[1], *Chi-Chun Lee*[1,2]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

ABSTRACT

Deception is an intended action of a deceiver to make an interrogator believe something is true (or false) that the deceiver believes to be false (or true) as a purposeful mechanism to share a mix of truthful and deceptive experiences when being asked to respond to questions. Conventionally, automatic deception detection from speech is regarded as a recognition task modeled only using the deceiver's acoustic cues and does not include temporal conversation dynamics between the interlocutors, i.e., ignoring the potential deception-related cues when the two interlocutors coordinate such a back-an-forth interaction. In this paper, we propose a joint learning framework to detect deception by simultaneously considering variations and patterns of the conversation using both interlocutor's acoustic features and their conversational temporal dynamics. Our proposed model achieves an unweighted average recall (UAR) of 74.71% on a recently collected Chinese deceptive corpus of dialog games. Further analyses reveal that the interrogator behaviors are correlated to the deceivers deception behaviors, and including the conversational features provides enhanced deception detection power.

Index Terms— deception, conversation, BLSTM, attention, speech acoustics

1. INTRODUCTION

Deception behavior is part of human nature in daily conversations and interactions. Investigating objective methods to accurately detect deceptive events in life has attracted attention especially among psychologists [1], law enforcement officers [2], and employers [3]. However, research has shown that human is not good at identifying deception [4], even for experienced experts such as police officers, prosecutors, and judges [5, 6], and the influence of personality factors in the deception detection ability has also been indicated [7]. Being a highly challenging task for humans, numerous research has examined automated detection approach based on a variety of measurable signal modality, e.g., word usage in text messenger [8], face thermal-imaging [9], brain's neuroimaging [10, 11, 12], electrodermal signals (EDA) [13], electroencephalography [14], and even keyboard stroke patterns [15].

Much of these past research rely on using specialized devices and does not scale easily for daily life applications. The most direct approach is to analyze verbal and non-verbal cues during human's natural communication [16]. In fact, with the rapid development of technical algorithms of machine learning to model speech, language, and video signals, emerging effort has largely concentrated on modeling these direct measurements of communicative behaviors for deception detection. For example, advancements have been observed in applications of fake news detection [17, 18], cyber crime detection [19], and even during employment interviews [20]. In this work, we focus on modeling speech-based cues for interlocutors during dyadic interactions. The verbal cues have also been shown to be more effective to detect deception than non-verbal cues, especially during dialog-based interviews [21].

Several works have recently developed algorithms for speech based deception detection in dialogs. Most of these works learn to detect deceptive events at either an "utterance-level", i.e., a ground truth label is given for every utterance, or a "question-level", i.e., a ground truth label is given for every unit of question. Some exemplary works include: Xie et al. developed a convolutional bidirection long short-term memory (CNN-BLSTM) for sentence-level deception detection using frame-level acoustic features as input [22]; Mendels et al. used six fully-connected layers deep neural networks (DNN) to detect sentence-level deception [23]; Levitan et al. analyzed a set of acoustic-prosodic features in differentiating truthful and deceptive responses to interview questions at question-level [24]. Furthermore, during interactions, deceptive behaviors not only are manifested in the acoustic characteristics but also are evident in the conversational turn-taking dynamics; for example, Vrij et al. showed that by examining conversational events, such as hesitations (e.g., "um", "hmmm"), pauses (silent) or latency periods (period of silence between consequent question and answer utterances), are helpful for untrained people identify liars [25].

Inspired by these works, we use a BLSTM based neural network to perform deception detection by simultaneously learning from deceiver's acoustic cues and conversational dynamics between the interrogator and the deceiver in a large Chinese corpus of dyadic game-based dialogs designed to study deception behaviors. Our proposed model, with its

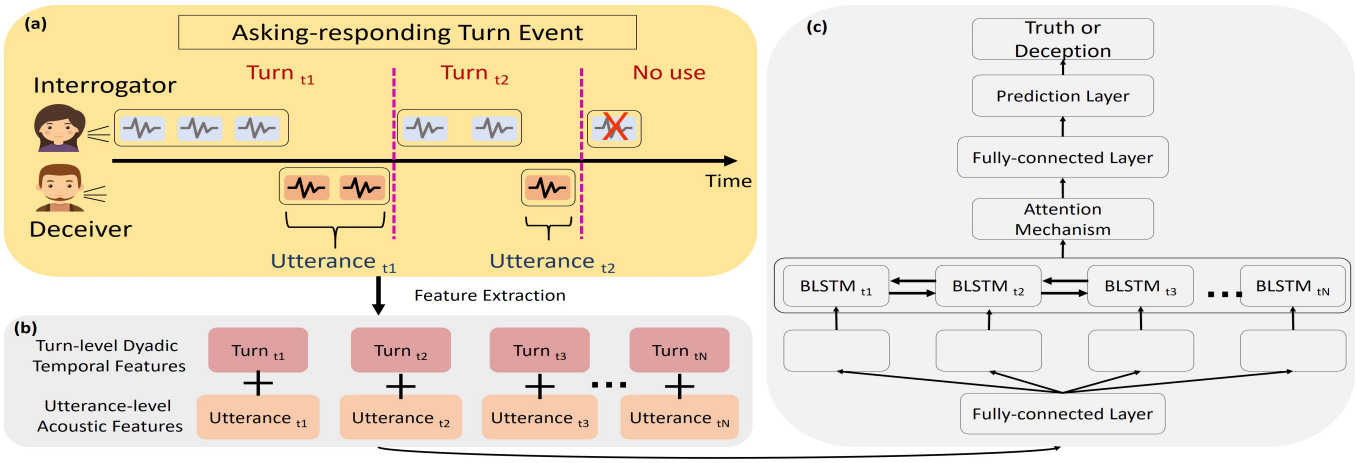

Fig. 1. (a) Turn segmentation (b) Feature-level fusion (c) Deception detection framework

inclusion of conversational temporal dynamics, obtains an truthful and deceptive classification accuracy of 74.71% unweighted average recall (UAR). We further provide analyses on the importance of these two different types of features used in indicating deception. The rest of paper is organized as follows: section 2 describes our database, methodology, and features design, section 3 includes experimental setup and results, and we finally conclude with future work.

2. RESEARCH METHODOLOGY

Fig.1 depicts our overall framework used in this work. The core idea is to model speech behaviors during interactions with conversational temporal and acoustic features. Temporal dynamics features contain conversational characteristics, and acoustic features describe deceivers speech acoustics. These features are then served as input to the detection network. The building block of the detection network is based on the structure proposed in [26], which consists of an initial fully-connected layer, then a bidirectional long short term memory (BLSTM) network with attention mechanism, and a final fully connected layers (BLSTM-DNN) for classification.

2.1. Daily Deceptive Dialogues Corpus of Mandarin

In this work, we use the Daily Deceptive Dialogues Corpus of Mandarin (DDDM) [27] recently collected at the National Tsing Hua University, Taiwan. It contains 27.2 hours of audio recordings of dyadic interactions from native speakers of Mandarin. This corpus includes 96 different speakers (48 male and 48 female) split in pairs into 48 interaction sessions; the subjects' age ranges from 20 to 25. There are a total of 7504 utterances in the database (segmented manually).

The database was collected using a protocol involving a pair of subjects playing games in a spontaneous conversational setting. One of the subjects played the role of an interrogator with the other player being the deceiver. The interrogator interviews the partner with topics chosen from a set of three daily activities, such as "have you ever attended any ball games or competed in ball games?", "have you ever attended/participated in any concerts?", or "have you ever attended/performed in any club achievement presentation?" with a goal to identify whether the deceiver was telling the truth with regard to each of the activities. Deceivers were instructed to deceive in their answers in at least one of the three topics discussed. Both sides of the subjects were provided with material incentive if they were capable to deceive effectively or identify the deceptive statements correctly.

In this study, we group segmented utterances into "ask-response" pairs showed in Fig.1(a) because the interrogator tends to ask questions attempting to identify whether the deceiver is telling the truth or not during the session. We use a complete "ask-response" pair as a time unit for our feature extraction, and within each pair, we can further categorize them as an asker-turn or a responder-turn (each turn may include multiple utterances from the same speaker). This segmentation method serves as the unit for inputting features into the BLSTM-based framework. This particular choice of unit is important as it indicates a complete *unit* that involves a connected context (i.e., one asking is linked to one responding, note that if an asking utterance has no related responses, we ignore those segments in this work). In summary, each topic is annotated by the deceiver indicating whether he/she is telling the truth or not, and each of this label includes multiple "ask-response" pairs.

2.2. Deception Detection Framework

Fig.1 shows our proposed detection architecture. Our deception detection model is built based on an BLSTM-DNN structure similar to a previous work [26]. In this work, our goal is to include both the conversational dynamics and the deceivers

Table 1. Results on the deception detection for the DDDM database (Aco. and Tem. means acoustic features and temporal features, respectively.)

Model	Human	SVM			DNN			LSTM-DNN			BLSTM-DNN		
Feature	-	Aco.	Tem.	Aco. + Tem.	Aco.	Tem.	Aco. + Tem.	Aco.	Tem.	Aco. + Tem.	Aco.	Tem.	Aco. + Tem.
UAR	55.55%	56.18%	49.12%	56.18%	70.62%	63.91%	71.20%	69.87%	64.19%	72.41%	70.31%	66.02%	**74.71%**
Deception	40.52%	56.12%	53.96%	54.68%	64.90%	71.20%	66.75%	72.00%	79.68%	70.83%	68.94%	77.91%	74.89%
Truth	70.59%	56.25%	44.44%	57.64%	76.34%	56.62%	75.65%	67.73%	48.69%	73.99%	71.67%	54.14%	74.53%
F1	54.71%	56.19%	49.00%	56.18%	69.65%	63.34%	70.68%	69.65%	62.75%	72.02%	70.03%	64.87%	74.39%
Precision	56.11%	56.18%	49.19%	56.16%	74.45%	64.87%	73.32%	70.25%	68.20%	74.03%	70.53%	68.37%	75.52%
Recall	55.55%	56.18%	49.20%	56.16%	70.62%	63.91%	71.20%	69.87%	64.19%	72.41%	70.31%	66.02%	74.71%

Model	Human	DT			LG			RF			AdaBoost		
Feature	-	Aco.	Tem.	Aco. + Tem.	Aco.	Tem.	Aco. + Tem.	Aco.	Tem.	Aco. + Tem.	Aco.	Tem.	Aco. + Tem.
UAR	55.55%	57.60%	54.77%	56.54%	55.83%	47.70%	56.89%	59.36%	54.06%	59.72%	60.78%	52.30%	61.48%
Deception	40.52%	58.99%	59.71%	56.12%	56.12%	48.20%	59.71%	58.99%	54.68%	57.55%	60.43%	69.78%	61.87%
Truth	70.59%	56.25%	50.00%	56.94%	55.56%	47.22%	54.17%	59.72%	53.47%	61.81%	61.11%	35.42%	61.11%
F1	54.71%	57.59%	54.67%	56.54%	55.83%	47.71%	56.86%	59.37%	54.07%	59.70%	60.78%	50.86%	61.49%
Precision	56.11%	57.62%	54.90%	56.53%	55.83%	47.71%	56.96%	59.36%	54.07%	59.70%	60.77%	52.95%	61.49%
Recall	55.55%	57.62%	54.86%	56.53%	55.84%	47.71%	56.94%	59.36%	54.07%	59.68%	60.77%	52.60%	61.49%

acoustic features as inputs to our detection network. We focus on deceivers acoustic cues and regard deceiver as the target speaker. The unit for deceiver's acoustic features is shown in the bottom of Fig.1 (a), which includes all of the utterances from the deceiver within a "ask-response" pair. The conversational temporal dynamic features are also computed within each of these "ask-response" pairs. In the following sections, we will describe in detail each of these features and our proposed use of the BLSTM classifier.

2.2.1. Utterance-level Acoustic Features

Previous research has shown that deception could be detected using a variety of prosodic features [24, 28]. In this work, we extract utterance-level acoustic features using the openSMILE toolbox [29] with the emobase config file. It contains 988 acoustic features per utterance. The emobase's low-level descriptors (LLDs) contains pitch (fundamental frequency), intensity (energy), loudness, cepstral (12 MFCC), probability of voicing, fundamental frequency envelope, 8 Line Spectral Frequencies (LSF), zero-crossing rate, and finally delta regression coefficients are computed from those LLDs. Then, the following functionals are applied to these extracted LLDs and their delta coefficients to generate the final 988 dimensional feature vector: maximum/minimum value and respective relative position within input, range, arithmetic mean (a mean), two linear regression coefficients and linear and quadratic error, standard deviation (std), skewness, kurtosis, quartile 1-3, and 3 inter-quartile ranges. They are further normalized to each speaker using z-score normalization.

2.2.2. Turn-level Conversational Temporal Features

The design of our conversational temporal dynamics features showed in Fig.2 is inspired by previous works on conversational analyses [25, 30, 31]. We design 20 dimensional temporal features based on conversational utterances in each "ask-response" pair. Interrogator and deceiver are first annotated with the role of "Ask" and "Res", respectively. For each of the asking/responding turn, we calculate the following features (all features are normalized to each speaker using z-score normalization):

- Duration: the total turn duration (d) of interrogators or deceivers utterances, denoted as Ask_d and Res_d.
- Duration difference: the duration difference between each of the interrogator and deceiver turn within a "ask-response" turn pair. It is calculated as $Res_d - Ask_d$ and $Ask_d - Res_d$.
- Duration addition: the sum of Res_d and Ask_d.
- Duration ratio: the ratio between Res_d and Ask_d, and Ask_d and Res_d.
- Utterance-duration ratio: the reciprocal ratio between the utterances length (u) and the turn duration (d) , denoted as Ask_{ud} and Ask_{du}, respectively.
- Silence-duration ratio: the reciprocal ratio between the silence (s) duration and the turn duration, denoted as Ask_{sd} and Ask_{ds}, respectively.
- Silence-utterance ratio: the reciprocal ratio between the silence duration and the utterances lengths, denoted by Ask_{su} and Ask_{us}, respectively.
- Hesitation time (h) (Response onset time): the difference between the onset time of the deceiver utterance and the offset time of the interrogator utterance, denoted as Res_h.

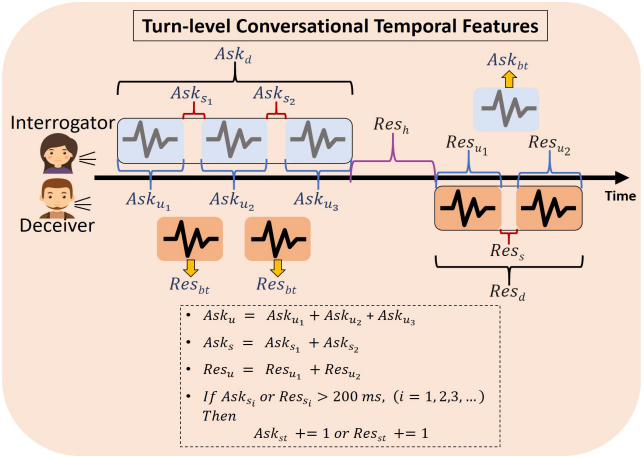

Fig. 2. This figure is the illustration diagram of turn-level conversational temporal features

- Backchannel times (bt): the number of time that a subject interrupts his/her interacting partner, denoted as Ask_{bt} and Res_{bt}.
- Silence times (st): the number of time that a subject produces a pause that is more than 200ms, denoted as Ask_{st} and Res_{st}.

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Experimental Setup

In this work, our basic building block network is BLSTM-DNN with attention. This model contains two fully-connected layers (dense layer) with Rectified Linear Unit (ReLU) [32] activation function, one BLSTM with attention layer, and finally one dense layer with softmax activation function. The number of hidden units is 16 in the first dense layer, 8 in BLSTMs with attention layer, 16 in the last dense layer. All layers include a dropout layer with 50% drop out rate.

We further compare our approach to a variety of baseline machine learning models. Specifically, these baseline models include Support Vector Machines (SVM) [13], ensemble method (AdaBoost and Random Forest (RF) [33]), linear model (Logistic Regression (LG) [34]), and non-parametric learning method (Decision Tree (DT) [35]). We also compare with other deep learning architecture that has been used for a similar task, such as feedforward neural network (DNN)[23, 28] and long-short time memory recurrent neural network (LSTM) [36] with attention mechanism.

Except LSTM model, the rest of the baseline models are static models. For those models, we compute 15 statistical functionals on each of the extracted turn-level acoustic/temporal dynamics features: maximum/minimum value and respective relative position within input, mean/median value, standard deviation, first percentile, ninety-ninth percentile, the difference between ninety-ninth percentile and first percentile, skewness, kurtosis, quartile 1, quartile 3, interquartile range. Each of the baseline model is trained after carrying out a uni-variate feature selection. The LSTM-DNN with attention model uses a similar architecture as our main BSLTM-DNN structure. LSTM with attention layer has 16 nodes and no bi-directional. In terms of the DNN baseline model, the framework is similar to a previous work [28], it consists of three fully-connected layers with ReLU activation function, and each layer includes a batch normalization layer [37] and dropout layer with 50% drop out rate. The number of hidden units are 16, 8, and 16, respectively.

We use 10 folds cross validation as our evaluation scheme with the metrics of unweighted average recall (UAR), F1 score, and precision. The BLSTM is trained with a fixed length (40 time-steps), which is the maximum length of turns in the DDDM corpus. We use zero-padding to make each data sample's time-steps the same if the length is less than 40 turns. In the training stage, the other hyperparameters, i.e., batch size and learning rate, is set to be 32 and 5*1e-4, respectively. These parameters are chosen with early stopping criteria in all conditions to minimize cross entropy on the validation set. The optimizer used in this work is ADAMMAX [38]. The whole framework is implemented using Pytorch toolkit [39]. For baseline methods, key hyperparameters, i.e., the number of estimators for Random Forest and AdaBoost, and the cost (C) of SVM, are grid searched within the range of [2, 4, 8, 16, 32, 64, 128, 256, 512], and [0.5, 0.1, 1 ,10], respectively.

3.2. Experimental Results and Analyses

Table 1 shows a summary of the complete recognition performances over different baseline methods. The BLSTM-DNN framework learned from the proposed temporal feature set with acoustic features obtains the best overall deception detection classification accuracy (74.71% UAR). This method surpasses methods with acoustic features-only, temporal features-only, and even human ability by 4.4%, 8.69%, and 19.16% absolute, respectively. The human accuracy is obtained by computing the concordance rates between interrogator's labeling of deceiver's topics versus deceiver's own labeling. Our results further demonstrate the importance in considering dyadic temporal conversational dynamics to improve the deception detection results. The other classifiers (only with the exception of the Decision Tree classifier) also reveal a similar finding.

One important observation is that when performing statistical t-tests between truthful and deceptive responses by the deceiver with respect to the conversational turn-taking features (shown in Table 3), conversational dynamic feature set obtained from the interrogators (i.e., Ask measures) behaviors play an important role in indicating whether the deceiver

Table 2. T-tests between truthful and deceptive responses in acoustic features (if a feature's p-value is smaller than 0.05 or 0.01, it is marked by O and *, respectively).

Feature	stddev	linregerrQ	linregerrA	range	max	iqr1-3	min	iqr1-2	quartile3	quartile1	iqr2-3	skewness	amean	linregc2	quartile2
$\Delta MFCC_{8th}$	O*	O*	O*	O	O	O	O	O	O	O	O				
$MFCC_{8th}$	O*	O*	O*	O			O*	O		O*		O	O		
$MFCC_{6th}$	O*	O*	O	O*	O										
$\Delta MFCC_{6th}$	O*	O*	O*	O											
$Loudness$	O	O	O*			O*					O*				
$\Delta MFCC_{12th}$				O	O										
$MFCC_{9th}$	O	O	O					O				O			
$\Delta F0_{Contour}$								O							
$\Delta MFCC_{7th}$									O		O				
ΔZCR														O	
$MFCC_{2th}$											O				
$\Delta VoicePro.$					O										
$\Delta Intensity$															O

Table 3. T-tests between truthful and deceptive responses in temporal features.

Tem.	P-value
Ask_{ud}	0.037
Ask_{su}	0.037
Ask_d/Res_d	0.041
Ask_{us}	0.043
Ask_{st}	0.052
Others	$>$0.2

is telling the truth or not. We further listen to the actual recordings and find that the interrogator would often ask more detailed questions and spend more time on thinking about what the next question they wanted to ask for times corresponding to when the deceivers were producing lies. This particular finding is quite intriguing as we observe that the "Human" labeled accuracy is relatively low on identifying deceptive events; however, their (interrogators) behaviors (may be unconsciously) would actually directly indicate whether he/she was indeed being given a truthful/deceptive answer.

Furthermore, in terms of the acoustic characterization, we observe a similar trend as previous works that the deceiver's acoustic manifestations do reflect whether their answers are truthful or deceptive (results shown in Table 2). There are 50 dimensions of acoustic parameters where p-values obtained are smaller than 0.05, and 16 features among them are smaller than 0.01. Specifically, $MFCC_{8th}$, $MFCC_{6th}$, and their first derivatives are useful for differentiating between truthful and deceptive responses for the deceivers, which is a similar result obtained compared with a previous work on an English dataset [24]. We also conduct a similar analysis on interrogators acoustic features; however, only two features (loudness and its first derivatives) obtain p-value that are smaller than 0.05.

4. CONCLUSIONS AND FUTURE WORK

The deceivers deception behavior cues exist not only in the deceiver's acoustic properties but also alter the interrogators asking behaviors as the two interlocutors engage in spontaneous game-based dialogs. In this work, we propose to design a set of dyadic conversational dynamics feature set that can be combined with the conventional acoustic feature sets to improve deception detection performances using a BLSTM-DNN with attention network architecture. Our method achieves a promising accuracy of 74.71% (UAR) on 2-class deception-truth recognition task. To the best of our knowledge, while there are many works in studying speech deception detection, this is one of the first works that have explicitly modeled the conversation dynamics together with the acoustic characteristics, and it further provides an analysis on the importance of different feature set in deception detection. In our immediate future work, we plan to extend our framework to include other behavior attributes, such as the lexical content to model the exact manner of question-answering content that may be indicative of a deceptive event. Furthermore, the variability of deception behaviors has been shown to be related closely to the deceivers personality [28, 7], a joint modeling of deceivers and interrogators personal attributes within a dynamic conversational setting may lead to further advancement in robust deception detection framework.

5. REFERENCES

[1] Erik Mac Giolla, Pär Anders Granhag, and Zarah Vernham, "Drawing-based deception detection techniques: A state-of-the-art review," *Crime Psychology Review*, vol. 3, no. 1, pp. 23–38, 2017.

[2] Timothy J Luke, Maria Hartwig, Emily Joseph, Laure Brimbal, Ginny Chan, Evan Dawson, Sarah Jordan, Pa-

tricia Donovan, and Pär Anders Granhag, "Training in the strategic use of evidence technique: Improving deception detection accuracy of american law enforcement officers," *Journal of Police and Criminal Psychology*, vol. 31, no. 4, pp. 270–278, 2016.

[3] S Rajkumar, "Assortment of uncertainty and randomness with fuzzy logic in deception detection for employee database management system using hotchpotch techniques," *World Appl. Sci. J*, vol. 21, no. 6, pp. 854–857, 2013.

[4] Maria Hartwig and Charles F Bond Jr, "Why do lie-catchers fail? a lens model meta-analysis of human lie judgments.," *Psychological bulletin*, vol. 137, no. 4, pp. 643, 2011.

[5] Leif Strömwall and Pär Anders Granhag, "How to detect deception? arresting the beliefs of police officers, prosecutors and judges," *Psychology, Crime and Law*, vol. 9, no. 1, pp. 19–36, 2003.

[6] Clea Wright and Jacqueline M Wheatcroft, "Police officers' beliefs about, and use of, cues to deception," *Journal of Investigative Psychology and Offender Profiling*, vol. 14, no. 3, pp. 307–319, 2017.

[7] Samuel D Spencer, "Examining personality factors in deception detection ability.," *Psi Chi Journal of Psychological Research*, vol. 22, no. 2, 2017.

[8] A Mbaziira and J Jones, "A text-based deception detection model for cybercrime," in *Int. Conf. Technol. Manag*, 2016.

[9] Ioannis Pavlidis, Norman L Eberhardt, and James A Levine, "Human behaviour: Seeing through the face of deception," *Nature*, vol. 415, no. 6867, pp. 35, 2002.

[10] Giorgio Ganis, "Deception detection using neuroimaging," *Detecting deception: Current challenges and cognitive approaches*, pp. 105–21, 2015.

[11] Tatia MC Lee, Mei-kei Leung, Tiffany MY Lee, Adrian Raine, and Chetwyn CH Chan, "I want to lie about not knowing you, but my precuneus refuses to cooperate," *Scientific reports*, vol. 3, pp. 1636, 2013.

[12] Mingming Zhang, Tao Liu, Matthew Pelowski, and Dongchuan Yu, "Gender difference in spontaneous deception: A hyperscanning study using functional near-infrared spectroscopy," *Scientific reports*, vol. 7, no. 1, pp. 7508, 2017.

[13] Jan Ondras and Hatice Gunes, "Detecting deception and suspicion in dyadic game interactions," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 2018, pp. 200–209.

[14] Yijun Xiong, junfeng Gao, and Ran Chen, "Connectivity network analysis of EEG signals for detecting deception," *Journal of Physics: Conference Series*, vol. 1176, pp. 032051, mar 2019.

[15] Merylin Monaro, Chiara Galante, Riccardo Spolaor, Qian Qian Li, Luciano Gamberini, Mauro Conti, and Giuseppe Sartori, "Covert lie detection using keyboard dynamics," *Scientific reports*, vol. 8, no. 1, pp. 1976, 2018.

[16] Aldert Vrij, "Deception and truth detection when analyzing nonverbal and verbal cues," *Applied Cognitive Psychology*, vol. 33, no. 2, pp. 160–167, 2019.

[17] Niall J Conroy, Victoria L Rubin, and Yimin Chen, "Automatic deception detection: Methods for finding fake news," *Proceedings of the Association for Information Science and Technology*, vol. 52, no. 1, pp. 1–4, 2015.

[18] Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu, "Unsupervised fake news detection on social media: A generative approach," in *Proceedings of 33rd AAAI Conference on Artificial Intelligence*, 2019.

[19] K. Veena and P. Visu, "Detection of cyber crime: An approach using the lie detection technique and methods to solve it," in *2016 International Conference on Information Communication and Embedded Systems (ICICES)*, Feb 2016, pp. 1–6.

[20] Marc-André Reinhard, Martin Scharmach, and Patrick Müller, "It's not what you are, it's what you know: Experience, beliefs, and the detection of deception in employment interviews," *Journal of Applied Social Psychology*, vol. 43, no. 3, pp. 467–479, 2013.

[21] Aldert Vrij, Sharon Leal, and Ronald P Fisher, "Verbal deception and the model statement as a lie detection tool," *Frontiers in psychiatry*, vol. 9, pp. 492, 2018.

[22] Y. Xie, R. Liang, H. Tao, Y. Zhu, and L. Zhao, "Convolutional bidirectional long short-term memory for deception detection with acoustic features," *IEEE Access*, vol. 6, pp. 76527–76534, 2018.

[23] Gideon Mendels, Sarah Ita Levitan, Kai-Zhan Lee, and Julia Hirschberg, "Hybrid acoustic-lexical deep learning approach for deception detection.," in *INTERSPEECH*, 2017, pp. 1472–1476.

[24] Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg, "Acoustic-prosodic indicators of deception and trust in interview dialogues," *Proc. Interspeech 2018*, pp. 416–420, 2018.

[25] Aldert Vrij, Maria Hartwig, and Pär Anders Granhag, "Reading lies: nonverbal communication and deception," *Annual review of psychology*, vol. 70, pp. 295–317, 2019.

[26] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.

[27] Chih-Hsiang Huang, Huang-Cheng Chou, Yi-Tong Wu, Chi-Chun Lee, and Yi-Wen Liu, "Acoustic indicators of deception in mandarin daily conversations recorded from an interactive game," in *Proceedings of the International Speech Communication Association (Interspeech)*, 2019, p. accepted for publication.

[28] Guozhen An, Sarah Ita Levitan, Julia Hirschberg, and Rivka Levitan, "Deep personality recognition for deception detection," in *Proc. Interspeech*, 2018, pp. 421–425.

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

[30] Aldert Vrij, Sharon Leal, Louise Jupe, and Adam Harvey, "Within-subjects verbal lie detection measures: A comparison between total detail and proportion of complications," *Legal and Criminological Psychology*, vol. 23, no. 2, pp. 265–279, 2018.

[31] Štefan Beňuš, Agustín Gravano, and Julia Hirschberg, "Pragmatic aspects of temporal accommodation in turn-taking," *Journal of Pragmatics*, vol. 43, no. 12, pp. 3001–3027, 2011.

[32] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[33] Sarah Ita Levitan, Yocheved Levitan, Guozhen An, Michelle Levine, Rivka Levitan, Andrew Rosenberg, and Julia Hirschberg, "Identifying individual differences in gender, ethnicity, and personality from dialogue for deception detection," in *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, 2016, pp. 40–44.

[34] Gangeshwar Krishnamurthy, Navonil Majumder, Soujanya Poria, and Erik Cambria, "A deep learning approach for multimodal deception detection," *arXiv preprint arXiv:1803.00344*, 2018.

[35] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo, "Deception detection using real-life trial data," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 59–66.

[36] Hamid Karimi, "Interpretable multimodal deception detection in videos," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 2018, pp. 511–515.

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

[38] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[39] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.