"*Your Behavior Makes Me Think It Is a Lie*": Recognizing Perceived Deception using Multimodal Data in Dialog Games

Huang-Cheng Chou[1,2] and Chi-Chun Lee[1,2]
[1]Department of Electrical Engineering, National Tsing Hua University
[2]MOST Joint Research Center for AI Technology and All Vista Healthcare
E-mail: hc.chou@gapp.nthu.edu.tw Tel: +886-3-5715131#34050
E-mail: cclee@ee.nthu.edu.tw Tel: +886-3-5162439

***Abstract*—Deception occurs frequently in our life. It is well-known that people are generally not good at detecting deception, however, behaviors of interlocutors during an interrogator-deceiver conversation may indicate whether the interrogator thinks the other person is telling deceptions or not. The ability to automatically recognize such a *perceived deception* using behavior cues has the potential in advancing technologies for improved deception prevention or enhanced persuasion skills. To investigate the feasibility to recognize the perceived deception from behaviors, we utilize a joint learning framework by considering acoustic-prosodic features, linguistic characteristics, language uses, and conversational temporal dynamics. We further incorporate personality attributes as an additional input to the recognition network. Our proposed model is evaluated on a Daily deceptive dialogue corpus of Mandarin database. We achieve an unweighted average recall of 86.70% and 84.89% on 2-class perceived deception-truth recognition tasks given the deceiver is telling either truths or lies, respectively. Further analyses unveil that 1) the deceiver's behaviors affect the interrogator's perception (e.g., the higher intensity of the deceiver makes the interrogator believe their statements even though they are deceptive in fact), 2) the interrogator's behavior features carry information about their own deception perception (e.g., interrogator's utterance duration is correlated to his/her perception of truth), and 3) personality traits indeed enhance perceived deception-truth recognition. Finally, we also demonstrate additional evidence indicating that human is bad at detecting deceptions – there are very few indicators that overlaps between perceived and produced truth-deceptive behaviors.**

I. Introduction

Deceptive behaviors often appear in our daily life. Despite its frequent occurrences, researchers have repeatedly shown that humans are not good at detecting deceptions, even for highly-skilled professionals, such as teachers, social workers, and police officers [1], [2] without advanced strategies (e.g, tactical use procedure [3], [4] or interview techniques that maximize deceivers cognitive load [5]) . Also, adults are no better at detecting children's lies than they are with adult lies [6]. Many studies have been carried out to investigate the potential reason underlying why humans perform so poorly at identifying deceptions (e.g., [7], [8], [9]). Due to the difficulty in identifying deception by humans, researchers have also developed an automatic deception detection system using different types of expressive facial modalities, such as facial action units [10], thermal facial analysis [11], [12], and facial expressions [13], [14]. Moreover, internal physiological measures [15] and even functional brain MRI [16], [17] have also been explored as potential bio-indicators of deception. While these bio-indicators can be useful in detecting deception, many of them require expensive and sometimes invasive instrumentation, which is not practical for real-world applications. Several recent works have demonstrated that speech and language cues carry substantial deceptive cues that can be modeled in automated deception detection tasks for potential large-scale deployment. For example, Chou et at. [18] indicated that the interlocutor's vocal characteristics and conversational dynamics should be jointly modeled to better perform deception detection in dialogs; Zhou et al. [19] found that language cues could be used to detect deceptions in computer-mediated communication messages, and Mbaziira et al. [20] developed linguistic-based deception detection for cybercrime.

Although these works have all worked on developing automatic deception detection systems, there are very few works attempting to understand why humans would *perceive* behaviors as deceptive or truthful regardless of whether it is a truthful or deceptive intent. Only recently, Chen et al. [21] tried to investigate the reasons why humans are poor at detecting deceptions by training classifiers to automatically recognize utterances that would be *perceived* either as truth or lies among those that were actual deceptive speech. In this paper, our work differs from [21] due to the following settings. Firstly, their perception ratings were derived from the third observer's point of view. To be more specifics, they recruited raters to re-annotate recordings. However, in our context, our ratings came directly from people of the interaction itself, i.e., corresponding to the real personal deceptive perception during the actual interaction. Additionally, Chen et al. [21] only focused on deceiver's features instead of both interlocutors (interrogator and deceiver) behavior cues. Lastly, most of the prior works treat the deception detection problem as an utterance-level classification task; however, this setting is unnatural since a human would alternate truthful and deceptive utterances to deceive enquirers. Hence, in this work, we follow

978-988-14768-8-3/20/$31.00 ©2020 APSIPA

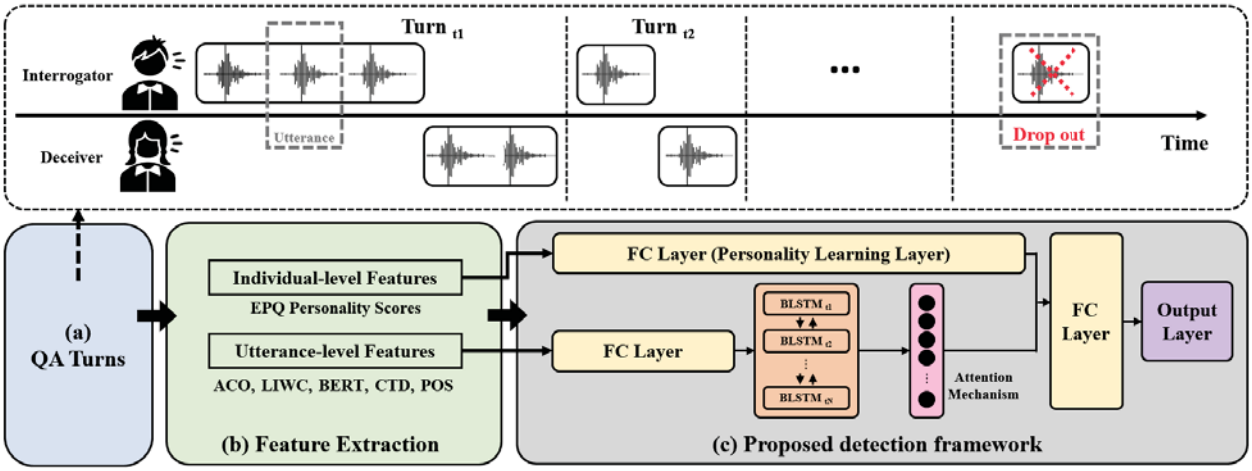

Fig. 1. (a) "Question-answering" (QA) Turns (b) Individual-level and utterance-level Feature Extraction (c) Proposed detection framework

our prior work [18] that models the whole conversations between interrogator and deceiver using "question-answering" (QA) turns events instead of single utterances.

In this paper, we use a Bidirectional Long Short-Term Memory (BLSTM) based deep neural network to recognize interrogator's *perceived* deception using multimodal cues, i.e., acoustic-prosodic features [22], Linguistic Inquiry and Word Count (LIWC) features [23], a lexical representation using a pre-trained BERT [24], and part-of-speech tags [25] in a recent Chinese corpus of Daily Deceptive Dialogue corpus of Mandarin (DDDM) that is designed specifically for the study of deception in dialogs [26]. Furthermore, while studies [27] have demonstrated that personality traits by itself may not serve as a direct predictor of deception, previous work has demonstrated that by fusing personality into the framework can be beneficial in developing an improved deception detection [28]. Hence, our proposed model integrates individuals' Eysenck Personality Questionnaire (EPQ) scores [29], [30], [31], [32], [33] as input features to further improve prediction accuracies of the *perceived* deception. Our proposed model that integrates the personality learning layer obtains an interrogator's *perceived* deception-truth classification accuracy of 86.70% and 84.89% unweighted average recall (UAR) when facing truths and deceptions tellers, respectively. Moreover, we show analyses on the importance of these five different types of features in revealing perceived prediction. The rest of paper is organized as follows: Section 2 explains the used corpus, proposed detection framework, and features extraction, Section 3 includes experimental setup and results, and we eventually conclude with future work.

II. RESEARCH METHODOLOGY

Fig. 1 illustrates the proposed structure used in this paper. Our main idea is to model speech and language behaviors during conversations by using acoustic-prosodic features, linguistic cues, conversational temporal features, and personality scores. Linguistic cues include language use, temporal dynamics features measure conversational characteristics, acoustic features describe speaker's vocal acoustics, and personality scores provide an assessment of characteristic patterns of thoughts, feelings, and behaviors. Afterward, these features are used as input to the detection network. The building block of the recognition system, BLSTM-DNN, is based on the structure proposed in [34], which contains an initial dense layer, then a bidirectional long short term memory (BLSTM) network with an attention mechanism, and a final dense layer for detection.

A. Daily Deceptive Dialogues Corpus of Mandarin

In this research, our proposed method is evaluated on the Daily Deceptive Dialogues Corpus of Mandarin (DDDM) [26]. It includes about 27.2 hours of audio-recordings of dyadic spontaneous interactions collected from native Mandarin speakers in Taiwan. This database contains 96 different speakers (48 males, 48 females), ranging from 20 to 25 years old, grouped in pairs into 48 interaction sessions. Participants were all asked to fill an Eysenck Personality Questionnaire (EPQ) [29] after they finished the experiment. There are a total of 7504 utterances segmented manually in the corpus.

DDDM was collected in a spontaneous conversational game setting, where each subject of a dyad took turns to play the role of an interrogator with the other player being the deceiver. The interrogator interviewed the deceiver with themes chosen from a set of questions about three daily activities: "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?". The intention of each interrogator was to recognize whether the deceiver (interlocutor) was telling the truth about each of question or not. To collect high-quality and real situation, deceivers were instructed to deceive in their answers in at least one of the three themes discussed.

TABLE I
RESULTS ON THE *perceived* DECEPTION-TRUTH DETECTION ON DDDM DATABASE (UAR (MACRO-RECALL), WEIGHTED-F1, MACRO-PRECISION) (%); INT., DEC. ARE DENOTED AS THE INTERROGATOR, AND THE DECEIVER, RESPECTIVELY.

Fea.	Condition		The Dec. is Telling Deceptions					The Dec. is Telling Truths				
	Who's Fea.	Who's EPQ	UAR	Deception	Truth	F1	Precision	UAR	Deception	Truth	F1	Precision
Aco.	Int.	-	69.10	61.03	77.16	70.88	68.21	68.98	54.48	83.49	70.57	63.59
		Int.	71.39	62.28	80.50	73.63	71.51	69.33	54.71	83.95	73.10	71.14
		Dec.	76.03	73.78	78.28	75.89	78.21	72.97	74.19	71.75	72.05	72.84
		Both	**76.52**	71.64	81.41	78.97	77.23	70.44	60.12	80.76	72.26	66.57
		Fun.	69.90	67.81	71.99	72.08	70.52	**73.91**	66.33	81.48	74.98	74.04
	Dec.	-	85.21	85.36	85.06	83.66	84.17	73.59	70.24	76.95	72.06	71.48
		Int.	84.63	88.42	80.85	82.68	81.21	77.00	70.26	83.73	77.34	78.47
		Dec.	86.46	87.72	85.20	85.44	84.91	76.10	77.26	74.95	75.50	72.46
		Both	**86.70**	86.97	86.43	86.07	86.75	**78.67**	80.02	77.31	78.45	77.54
		Fun.	82.88	88.97	76.80	79.91	82.80	73.79	66.60	80.99	75.68	72.93
	Both	-	79.88	76.22	83.55	78.27	79.01	65.57	50.05	81.09	65.70	64.17
		Int.	81.45	84.75	78.16	80.25	78.50	67.47	59.60	75.35	68.53	62.54
		Dec.	79.65	84.17	75.14	76.87	77.75	71.16	58.12	84.19	70.85	76.44
		Both	**83.93**	88.75	79.11	82.09	82.06	70.22	59.02	81.41	73.58	70.02
		Fun.	80.79	80.36	81.22	80.30	79.36	**75.40**	65.50	85.31	77.12	77.49
CTD	Both	-	65.73	56.53	74.94	67.51	67.09	75.20	92.64	57.75	70.58	73.84
		Int.	76.26	76.78	75.74	77.34	74.83	73.57	83.12	64.01	71.00	69.01
		Dec.	**78.04**	82.72	73.36	78.78	77.05	**77.37**	86.10	68.64	74.21	72.29
		Both	74.12	62.25	85.99	75.49	75.81	71.83	73.52	70.13	72.13	68.78
		Fun.	63.34	63.58	63.10	62.95	64.40	71.64	69.10	74.18	72.10	68.18
BERT	Int.	-	72.26	82.03	62.49	69.62	76.14	68.28	64.12	72.43	69.83	67.94
		Int.	74.69	77.08	72.29	75.20	76.14	71.33	63.29	79.38	74.12	72.07
		Dec.	73.54	73.36	73.72	74.84	72.81	69.75	65.88	73.63	71.29	74.91
		Both	**80.26**	83.22	77.29	80.56	81.11	72.71	62.98	82.43	75.68	72.04
		Fun.	64.14	54.56	73.73	67.95	63.04	**74.46**	69.45	79.47	76.57	76.21
	Dec.	-	63.75	49.14	78.36	68.98	62.22	70.07	76.79	63.36	68.09	66.92
		Int.	72.23	68.00	76.47	71.15	74.31	72.90	78.98	66.81	72.33	71.78
		Dec.	74.91	76.28	73.55	75.63	73.12	74.54	74.43	74.65	74.44	71.27
		Both	**75.95**	75.67	76.23	73.52	74.69	**74.87**	75.26	74.48	74.14	71.18
		Fun.	65.92	71.14	60.70	66.49	62.78	73.71	67.98	79.45	71.96	73.55
	Both	-	71.41	70.33	72.49	74.12	72.93	62.84	50.76	74.92	69.63	66.20
		Int.	75.03	74.39	75.67	75.99	74.94	73.83	65.05	82.61	76.94	73.41
		Dec.	**78.38**	75.25	81.50	80.23	81.18	73.82	74.87	72.78	74.80	72.34
		Both	76.32	80.78	71.86	76.43	74.93	70.63	61.31	79.95	72.34	70.77
		Fun.	65.91	77.39	54.43	65.99	66.21	**73.91**	73.05	74.76	73.93	72.91

Both sides of the participants were supplied with material incentive if they deceit effectively or identify the truthful and the deceptive statements correctly.

In this paper, we followed [18] to split segmented utterances into "question-answering" (QA) pairs illustrated in Fig. 1 (a). Since the interrogator tends to ask questions to facilitate detecting whether the deceiver's were being truthful or deceptive during the conversation, we use a complete QA pair as a time unit for our feature extractions. Within each pair, we can further group them as a question-turn or an answering-turn. Note that each turn may consist of multiple utterances from the same speaker. This segmentation method in [18] serves as the unit to input multimodal features into the BLSTM-DNN structure. This special choice of unit is critical as it reveals a whole unit that includes a linked context (i.e., one question is bind to one answering, note that if a question utterance has no associated responses, we disregard those segments in this study illustrated in the top of Fig. 1). To be noticed, each topic is not only annotated by the interrogator but also by the deceiver. Therefore, we know both which topic the deceiver is telling the truth or not and also the interrogators perception about deception for each topic. Each of these labels includes multiple QA pairs. In summary, we utilize the labels from the interrogator to train our detection model for predicting perceived behavior of the interrogator.

B. Perceived Deception-Truth Detection Framework

Fig. 1 describes our proposed detection model. Our main component is a BLSTM-DNN structure similar to a previous [34]. In this study, our target is to contain both speaker's personality information and multimodal speech and language behavior features as inputs to our detection network. To be noticed, this study differs from [18], which only pay attention to deceivers' acoustic-prosodic cues and consider deceivers as the target speaker. Specifically, we regard the interrogator as our target speaker in this paper. The unit for interrogators' and deceivers' acoustic-prosodic features is displayed in the top of Fig. 1 (a), which incorporate all of the utterances from the interrogator and the deceiver within a "question-answering" (QA) pair. The following features are computed within each of these QA pairs. The rest of sections, we will describe in detail each feature and the proposed use of the BLSTM-based classifier.

1) Utterance-level Acoustic-Prosodic Features: Many studies have shown that a variety of prosodic and acoustic features

TABLE II
RESULTS ON THE *perceived* DECEPTION-TRUTH DETECTION ON DDDM DATABASE (UAR (MACRO-RECALL), WEIGHTED-F1, MACRO-PRECISION) (%); INT., DEC., AND BOTH ARE DENOTED AS THE INTERROGATOR, THE DECEIVER, AND BOTH INTERLOCUTORS, RESPECTIVELY.

Fea.	Condition		The Dec. is Telling Deceptions					The Dec. is Telling Truths				
	Who's Fea.	Who's EPQ	UAR	Deception	Truth	F1	Precision	UAR	Deception	Truth	F1	Precision
POS	Int.	-	76.22	66.42	86.03	78.18	77.11	72.92	75.43	70.41	71.29	69.08
		Int.	**81.04**	81.42	80.67	80.79	78.95	78.32	80.29	76.36	76.92	75.23
		Dec.	74.75	64.50	85.00	75.83	75.66	78.40	82.07	74.74	78.27	76.88
		Both	80.11	73.64	86.58	81.15	79.93	**80.04**	77.71	82.37	80.26	77.36
		Fun.	69.05	60.44	77.65	70.71	70.39	79.28	74.55	84.01	81.06	78.83
	Dec.	-	71.92	72.67	71.18	71.61	70.37	73.05	78.48	67.63	71.88	69.90
		Int.	72.16	67.81	76.52	74.69	71.27	**81.41**	87.38	75.44	79.04	76.93
		Dec.	71.46	69.36	73.57	74.01	71.85	78.10	76.21	79.99	81.09	75.65
		Both	**73.94**	67.89	79.99	74.49	73.86	79.85	80.38	79.31	79.15	76.11
		Fun.	64.56	63.03	66.09	67.26	66.53	77.49	72.26	82.72	77.96	77.71
	Both	-	77.39	77.17	77.60	76.62	75.74	72.61	64.69	80.53	70.66	67.91
		Int.	**81.19**	85.44	76.93	77.96	78.81	79.72	81.79	77.65	78.86	75.34
		Dec.	78.54	74.89	82.19	78.84	77.02	**84.89**	88.21	81.56	82.98	81.12
		Both	79.70	77.06	82.34	79.11	79.43	81.45	84.79	78.12	80.47	77.07
		Fun.	71.10	73.56	68.64	72.80	70.61	79.20	75.76	82.64	79.72	77.09
LIWC	Int.	-	72.04	61.56	82.53	73.65	72.41	74.26	63.83	84.68	72.88	71.91
		Int.	78.03	78.92	77.14	77.49	77.47	72.29	74.60	69.98	70.36	70.00
		Dec.	77.27	74.06	80.49	76.01	75.61	**80.47**	83.71	77.23	80.03	76.54
		Both	**80.13**	85.22	75.03	77.16	77.59	74.90	72.02	77.78	74.54	70.49
		Fun.	70.68	72.11	69.25	71.80	70.60	74.58	70.74	78.43	74.34	73.97
	Dec.	-	68.17	69.83	66.50	68.70	66.16	59.18	49.07	69.29	64.61	57.47
		Int.	73.40	73.36	73.43	73.96	73.59	**74.53**	72.55	76.51	74.84	71.78
		Dec.	71.50	72.94	70.05	72.69	70.72	73.77	77.90	69.64	73.50	72.64
		Both	**73.72**	75.67	71.77	72.18	71.44	69.89	63.86	75.93	70.93	67.05
		Fun.	67.04	69.89	64.20	64.97	67.00	70.16	69.43	70.90	69.64	67.31
	Both	-	67.78	60.47	75.09	64.98	64.46	70.52	71.93	69.12	70.27	66.58
		Int.	73.18	70.94	75.41	73.01	72.91	69.62	68.62	70.62	70.94	67.54
		Dec.	71.12	73.31	68.94	71.35	70.02	74.76	69.60	79.92	74.74	74.39
		Both	**75.05**	72.97	77.12	74.83	72.85	**75.56**	79.36	71.76	74.30	73.23
		Fun.	70.66	80.78	60.54	64.45	70.05	72.71	67.67	77.75	73.69	71.69

could be useful indicators of deception, such as pitch, Mel-Frequency Cepstral Coefficients (MFCC), and intensity [28], [35], [18], [21]. In this study, we extract a similar set of utterance-level acoustic-prosodic features using the openSMILE feature extraction toolkit with the emobase config file [22]. It includes 988 acoustic-prosodic features per utterance. To be more specific, the emobases low-level descriptors (LLDs) contains pitch (fundamental frequency, F0), intensity (energy), loudness, cepstral (12 MFCC), probability of voicing (VoicePro), fundamental frequency envelope, 8 Line Spectral Frequencies (LspFreq), zero-crossing rate (ZCR), and finally delta regression coefficients are computed from those LLDs. Then, the functionals[1] are applied to these extracted LLDs and their delta coefficients to generate the final 988-dimension feature vector. The more detailed information is in [22]. They are further normalized to each speaker using z-score normalization and denoted by "ACO" in the TABLE I.

2) Turn-level Conversational Temporal Dynamics: This feature set was first proposed in [18] which was inspired by past studies on conversational analysis [36], [37], [38]. It contains 20-dimensional temporal features based on conversational utterances in each QA pair, such as silence-duration ratio, utterance-duration ratio, silence-utterance ratio, backchannel times, etc. All features are normalized to each speaker using z-score normalization and denoted as "CTD". A brief description of these feature extractions are in [18]. The interrogator and the deceiver are first annotated with the role of "Ask" and "Res" in [18]. Then, for each of the asking/responding turn, we calculate the features, but we only mention the following features showed in TABLE III and TABLE VI:

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 utterance lengths, denoted by Ask_{su} and Ask_{us}, respectively.
Backchannel times (*bt*): the number of times that a subject interrupts his/her interacting partner, denoted as Ask_{bt} and Res_{bt}.

3) Utterance-level Textual Representation: In order to consider language use as features in the task of perceived deception-truth detection, we recruited three university students from the Department of Chinese Literature to transcribe the audio recordings. We perform word segmentation for each utterance using the CKIP-style Chinese Natural Language Processing (NLP) tools [25]. In this work, we use BERT [24] as model for deriving the textual representation. Specifically, we use BERT-Base in Chinese version and extract 768-dimension

sentence encoding using the utterances in each QA pair. All features are also normalized to each speaker using z-score normalization, denoted by "BERT".

4) Utterance-level Linguistic Inquiry and Word Count Feature Extraction: Previous research [39], [40] has used different word usage patterns to train a deception detection model specifically with the features derived from the Linguistic Inquiry and Word Count (LIWC) [23]. Furthermore, Levitan et al. [41] investigated the perception of deception and identified characteristics of statements that are perceived as truthful or deceptive by interviewers based on LIWC features. Inspired by these studies, we extracted a total of 83-dimensional features using LIWC 2015 in this work after performing word segmentation preprocessed by CKIP-style Chinese NLP tools [25]. That is, these features includes standard linguistic dimensions (e.g., negations, impersonal pronouns, 2^{nd} person, auxiliary verbs), markers of psychological processes (e.g., affective (negative emotion), social, cognitive (discrepancy), perceptual, biological processes, drivers (like power, number, risk), relativity (e.g., motion, time, space), personal concerns (e.g., work, leisure, death), and informal language (e.g., swear, assent, non-fluences words).

In the following, we clarify our terminology with a few examples for the analysis presented in Section III-B:

Linguistic Dimensions: $Negate$ (Negations, e.g., no, not, never), $Ipron$ (Impersonal pronouns, e.g., it, its, those), You (2nd person, e.g., you, your, thou), $Auxverb$ (Auxiliary verbs, e.g., am, will, have)

Psychological Processes: $Negemo$ (negative emotion, e.g., hurt, ugly, nasty)

Cognitive Processes: $Discrep$ (discrepancy, e.g., should, would)

Other Grammar: $Compare$ (comparisons, e.g., greater, best, after)

Drivers:: $Power$ (power, e.g., superior, bully), Number (number, e.g., second, thousand)

Informal Language: $Nonflu$ (non-fluencies, e.g., er, hm, umm)

5) Utterance-level Part-of-speech Tagging Feature Extraction: Researchers have also analyzed part-of-speech tagging (POS), using toolkits such as NLTK [42], Stanford Parser[43], and CKIP parser [44]) to train the deception detection model [45], [46], [47], [48]. All of these studies were conducted on English database. In this work, we extracted 48-dimensions POS tags using the CKIP-style Chinese NLP tools [25]. Grammatical and syntactical structure of each utterance (e.g., non-predictive adjective, coordinate conjunction, adverb of degree, verb, quantitative) within a QA pair is captured through analzing the transcripts. Computed features demonstrate the distribution of these categories in percentage for each utterance. Then, all features are normalized to each speaker using z-score normalization. To clarify the notation used in Section III-B, we denote the following POS features (the more detailed explanations of these features are in [25]):

A: the adjective
DM: the measure (number)
Dk: the sentential adverb
Dfb: the adverb after verbs
Na: the common noun
Neqb: the determinatives after the measure
Nv: the nominalization verbs
VI: the state intransitive verb
VCL: the transitive verb before the place

6) Eysenck Personality Questionnaire Personality Traits: All participant in the DDDM database had been asked to fill out Eysenck Personality Questionnaire (EPQ), which characterizes the personality traits of a person using the following four factors:

Extraversion (E): People whose scores are high in this dimension show the characteristics of outgoing, talkative and desire to explore. People whose scores are low tend to have a stable emotion, to stay distant to people except intimate individuals, and to lead a regular life. (Adjective words: sociable, active, sensation-seeking)

Neuroticism (N): It is characterized as a normal behavior instead of symptoms. High scores might imply depression and anxiety so as to lack of rationality. (Adjective words: moody, lack of autonomy, low self-esteem)

Psychoticism (P): It exists in all individuals with different degrees. People whose scores are high might love to stay lonely and to be in consideration so as to be difficult to adapt to a new environment. (Adjective words: tough-minded, masculine, manipulative)

Lie (L): L is a validity scale which assesses the person's tenderness for lying or pretending good.

To be used in practice, the zscore normalization of the EPQ score function is computed in the training data, and then apply it on the test data. Furthermore, inspired by [49], [50] that computed statistics (e.g., mean, maximum, minimum) as measures of personalities for each interaction unit, e.g., within a group, we not only include raw EPQ scores but also compute seven statistics (difference, maximum, minimum, mean, standard deviation, lower quartile (quartile1), and upper quartile (quartile3)) between interrogator and deceiver (each pair participant).

III. Experimental Setup And Results

A. Experimental Setup

The proposed models in this paper are based on BLSTM-DNN with attention mechanism containing one fully-connected layer (dense layer) with ReLU activation function, one BLSTM layer with an attention mechanism, one dense layer with ReLU activation function, and finally one dense layer with softmax activation function (prediction layer). Additionally, we add one dense layer (personality learning layer) with the ReLU activation function for a joint training with personality information. The number of hidden units is 16 in the first dense layer and in the last dense layer; 8 is in BLSTM layer with an attention mechanism. Also, a dropout layer is added for these layers excluding personality learning and prediction layers with a 50% dropout rate. Moreover, the number of hidden units of the additional dense layer is

TABLE III
T-TESTS BETWEEN PERCEIVED TRUTHFUL AND DECEPTIVE RESPONSES IN FIVE DIFFERENT TYPES OF FEATURES, AND A FEATURE'S P-VALUE ALL IS SMALLER THAN 0.05 (IF A FEATURE'S P-VALUE IS SMALLER THAN 0.01, IT IS MARKED BY *). NOTE THAT ONLY TWO FEATURES' P-VALUE ARE SLIGHTLY HIGHER THAN 0.05, THEIR P-VALUE IS AROUND 0.067 MARKED BY -. THE DETAILED INFORMATION ABOUT ACOUSTIC-PROSODIC FEATURES ARE IN TABLE IV AND TABLE V.

Conditions	Feature Set	Interrogator's Features	Deceiver's Features
The Dec. Is Telling Truths	Aco.	$MFCC_{1*,2*,3,9,11,12th}$, $\Delta MFCC_{6,11th}$ $LSP_{0,7th}$, $Loudness$*, ZCR $VoicePro$, $\Delta VoicePro$	$\Delta MFCC_{0,2,4,6,8}$ $LSP_{0,6,7th}$, $Loudness$, $Intensity$ ΔZCR, $\Delta F0$
	CTD	Ask_{sd}*, Ask_{ud}*, Ask_{us}, Ask_{su}, Ask_{bt}	-
	BERT	88 features (1 of them*)	80 features (2 of them*)
	POS	DM*, Dfb*, VCL*	DM*, Dfb, DK
	LIWC	$Compare$, $Power$	$Number$*, $Auxverb$,$Discrep$,$Nonflu$
	EPQ	-	-
	Fun. EPQ	-	
The Dec. Is Telling Deceptions	Aco.	$MFCC_{2*,4,5,6,7,8,9th}$, $\Delta MFCC_{1,2,3,4,7,8,9,11,12th}$ $\Delta LSP_{0,1,2,4,6th}$, $Loudness$*, ΔZCR $Intensity$*, $\Delta Intensity$, $VoicePro$, $\Delta VoicePro$	$MFCC_{1,3,6,7,11th}$, $\Delta MFCC_{4,5,7,8,9,11}$ $LSP_{0,1,2,6,7th}$, ΔLSP_{6th}, ZCR, $Loudness$, $\Delta Loudness$ $Intensity$*, $\Delta Intensity$, $\Delta VoicePro$, $\Delta F0$, $\Delta F0env$
	CTD	-	Res_{ud}, Res_{sd}, Res_{us}
	BERT	104 features (6 of them *)	57 features (3 of them *)
	POS	DM*, $Neqb$*, VI*, Na	A*, DK, Nv
	LIWC	$Negate$*, $Ipron$	You, $negemo$, $compare$
	EPQ	L	-
	Fun. EPQ	$L^{-}_{Difference}$, $E^{-}_{Quartile3}$	

equal to the size of personality scores input features. That is, since every subject has 4-dimensional personality scores, and personality information inputs have 4, 8, and 28 hidden units when using a single speaker, both speakers, or both speakers with seven statistics, respectively.

All experiments are evaluated using 10-fold cross-validation scheme with the metric of unweighted average recall (UAR, which is equal to macro-recall), macro-precision, and weighted-F1. The BLSTM layer is trained with a fixed length (40 time-steps), which is the maximum length of turns in the used dataset. To be noticed, we use zero-padding to make all samples 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 0.005, respectively. These parameters are chosen with early stopping criteria in all conditions to minimize cross-entropy loss on the validation set. The optimizer used in this work is ADAMMAX [51], and our implementation is based on PyTorch toolkit [52].

B. Experimental Results and Analyses

TABLE I presents a summary of the complete results in recognizing *perceived deception* under two conditions (the deceiver is telling either truths or lies), and the *Int.*, *Dec.*, and *Both* means the interrogator, the deceiver, and both interlocutors, respectively. The column, *Who's Fea.*, in TABLE I and II implies that the feature comes from whom, such as the interrogator, the deceiver, or both interlocutors. Besides, the column, *Who's EPQ.*, suggests that the EPQ scores are from the interrogator, the deceiver, both interlocutors, or the both values applied with seven statistics. Moreover, our proposed framework learned from both of the interrogators and the deceiver's acoustic-prosodic features and their personality traits obtains the best overall *perceived* deception-truth recognition tasks *when the deceiver is telling deceptions* (86.70% UAR). On the other hand, *when the deceiver is telling truths*, the proposed model learned from the deceiver's EPQ personality scores and part-of-speech taggers (POS) achieves the best performance (84.89% UAR). Our designed method surpasses methods with the same features but without any speaker's personality information by 1.49%, and 12.28% absolute when the deceiver is telling deceptions and truths, respectively. Our results demonstrate the importance in considering the personality traits to improve the perceived truths and deceptions detection results. Most of our results reveal a similar pattern, and we also find the performance benefits from modeling both the interrogator and the deceiver. For instance, when the deceiver is telling the truths, the model learned from the interrogator's LIWC features with the deceivers personality information performs better. It seems that there is a complementary information between the deceivers personality scores and the interrogator's behavior features.

Another observation is that when performing statistical t-test between *perceived* truthful and deceptive responses by the interrogator with respect to the QA pairs (exhibited in Table III), acoustic-prosodic feature set obtained from the interrogators behaviors play a significant role in showing whether the interrogator perceived the deceiver is telling the truth or not. On the other hand, the deceiver's acoustic-prosodic features have more significant influence on the *perceived* deception-truth detection of the questioners while the deceiver is telling the deceptions. The more detailed descriptions are in Section II-B1. In addition, interestingly, we found that the length of utterance duration from the deceivers have higher correlation with the perceived truthful prediction when the deceiver is telling deceptions. On the contrary, the interrogators tends to predict the truth when their length of utterance duration is longer. Furthermore, we notice that there are more significant dimensions from interrogators than deceivers when examining the results obtained using BERT features.

In terms of LIWC features (the details are in Section II-B5), the results pointed out different indicators in language use when performing recognition under different conditions. According to [53], [54], people might use more negative words when lying, and our results obtain the same trends. We

TABLE IV
T-TESTS BETWEEN PERCEIVED TRUTHFUL AND DECEPTIVE RESPONSES IN ACOUSTIC-PROSODIC FEATURES WHEN THE DECEIVER IS TELLING DECEPTIONS, AND A FEATURES VALUE ALL IS SMALLER THAN 0.05 (IF A FEATURE'S P-VALUE IS SMALLER THAN 0.01, IT IS MARKED BY *) 19 STATISTICS DENOTE BY [1]

Who's Fea.	Int.	Dec.
$\Delta F0$	-	2*, 3*, 4, 5, 8*, 9, 14*, 16, 19
$\Delta F0env$	-	10*, 18, 19
$Intensity$	2	10, 13, 17
$\Delta Intensity$	2, 3, 7, 14	5, 12, 17
$Loudness$	2*, 15	10, 17, 19
$\Delta loudness$	2, 3, 4, 8, 14, 16	2, 5, 14
$LspFreq_{0th}$	-	2, 3*, 4*, 16*
$LspFreq_{1th}$	-	1, 7, 13, 14*, 15, 16
$LspFreq_{2th}$	-	1, 14, 15, 16, 18
$LspFreq_{6th}$	-	7*
$LspFreq_{7th}$	-	2, 3, 8, 9, 19
$\Delta LspFreq_{0th}$	15	-
$\Delta LspFreq_{1th}$	18	-
$\Delta LspFreq_{2th}$	2	-
$\Delta LspFreq_{4th}$	4	-
$\Delta LspFreq_{6th}$	2, 3, 5	4, 8, 9, 16
$MFCC_{1th}$	-	1, 4, 15, 16
$MFCC_{2th}$	5	-
$MFCC_{3th}$	-	3, 8, 9*, 19
$MFCC_{4th}$	5, 13, 18	-
$MFCC_{5th}$	3, 4*, 6, 8, 9	-
$MFCC_{6th}$	2, 13, 18	7
$MFCC_{7th}$	18	7
$MFCC_{8th}$	12	-
$MFCC_{9th}$	3, 4*, 14, 19	-
$MFCC_{11th}$	-	1, 2*, 7*, 15, 16
$\Delta MFCC_{1th}$	13	-
$\Delta MFCC_{2th}$	5	-
$\Delta MFCC_{3th}$	18	-
$\Delta MFCC_{4th}$	2, 18	1
$\Delta MFCC_{5th}$	-	5
$\Delta MFCC_{7th}$	7	16
$\Delta MFCC_{8th}$	7, 8, 9, 19	15
$\Delta MFCC_{9th}$	3, 4, 7, 14, 16	6
$\Delta MFCC_{11th}$	14	6, 7
$\Delta MFCC_{12th}$	6, 7	-
$VoiceProb$	13	-
$\Delta VoiceProb$	13*	3, 4, 8*, 9, 19*
ZCR	-	1, 2*, 3, 15*, 16
ΔZCR	11	-

TABLE V
T-TESTS BETWEEN PERCEIVED TRUTHFUL AND DECEPTIVE RESPONSES IN ACOUSTIC-PROSODIC FEATURES WHEN THE DECEIVER IS TELLING TRUTHS, AND A FEATURES VALUE ALL IS SMALLER THAN 0.05 (IF A FEATURE'S P-VALUE IS SMALLER THAN 0.01, IT IS MARKED BY *) 19 STATISTICS DENOTE BY [1].

Who's Fea.	Int.	Dec.
$\Delta F0env$	-	6
$Intensity$	-	3
$Loudness$	5, 12, 18	3, 16
$LspFreq_{0th}$	1*, 14, 15*, 16	-
$LspFreq_{7th}$	16	13
$\Delta LspFreq_{0th}$	-	5
$\Delta LspFreq_{6th}$	-	6
$MFCC_{1th}$	16	-
$MFCC_{2th}$	1*, 7, 14*, 15*, 16*	-
$MFCC_{3th}$	1*, 14, 15, 16	-
$MFCC_{9th}$	1, 14	-
$MFCC_{11th}$	1, 15	-
$MFCC_{12th}$	14, 15	-
$\Delta MFCC_{2th}$	-	1
$\Delta MFCC_{4th}$	-	6
$\Delta MFCC_{6th}$	2	18
$\Delta MFCC_{8th}$	-	5
$\Delta MFCC_{11th}$	7	-
$VoiceProb$	9, 19*	-
$\Delta VoiceProb$	8	-
ZCR	1, 2, 6	-
ΔZCR	-	1

also find that the interrogator tends to perceive the deceiver's statements as deceptions when the interrogator uses more negations, and those statements indeed are deceptive. Besides, we can also use the patterns of grammar structure (POS) to detection the interrogator's perceived predictions. In our observations, the interrogator turns to predict the statement as deceptions when the deceiver uses more the adjective and the sentential adverbs (Dk, e.g., anyway or it is said).

The most interesting part is about *L* (one factor in EPQ), it is related to the person's tendency for lying or pretending good. The *L* might be an important factor in telling the interrogator's perceived prediction. In addition, the difference in *L* dimension between the interrogator and the deceiver also could be a key indicator even though its p-value is slightly higher than 0.05.

Further, we observe a similar trend as previous research manifested on the English database [21] that the interrogator judged high intensity utterances as truths because the louder utterances might be perceived as more confident even though these utterances could be deceptive in fact. Besides, when the deceiver is telling the truths, there are 7 dimensions of the deceivers acoustic-prosodic parameters where p-values obtained are small than 0.01, and there are 5 features among them that are smaller than 0.05. On the other hand, when the deceiver is telling the deceptions, there are 16 dimensions of the interrogator's features where p-values are small than 0.01, and there are 43 acoustic-prosodic features among them are smaller than 0.05.

Lastly, we perform further analyses showed in TABLE VI by examining the intersections between Table III and t-tests between produced truthful and deceptive responses across four different types of features and indicate those features' p-value that is smaller than 0.05. There are only very few acoustic-prosodic features left, and it might explain the reason why the human is bad at detecting deception from speech acoustics directly. However, surprisingly, the conversational dynamics features from inquirers are useful to be the indicators of detecting both produced and perceived deceptions and truths. Last but not at least, one of part-of-speech tagging from interrogators, DM, is also an important factor in predicting perceived truth-deception.

IV. CONCLUSIONS AND FUTURE WORK

In this study, the proposed framework are used in automatic detecting how humans *perceive* truths and deceptions when the interlocutor is telling the truths or deceptions. We analyzed a full suite of acoustic-prosodic features, linguistic cues,

[1](1): amean, (2): iqr1-2, (3): iqr1-3 , (4): iqr2-3, (5): kurtosis, (6): linregc1, (7): linregc2, (8): linregerrA, (9): linregerrQ, (10): max, (11): maxPos, (12): min, (13): minPos, (14): quartile1, (15): quartile2, (16): quartile3, (17): range, (18): kewness, (19): stddev

TABLE VI
ALL FEATURES ARE THE INTERSECTIONS BETWEEN TABLE III AND T-TESTS BETWEEN PRODUCED TRUTHFUL AND DECEPTIVE RESPONSES IN FOUR DIFFERENT TYPES OF FEATURES, AND A FEATURE'S P-VALUE ALL IS SMALLER THAN 0.05.

Conditions	Feature Set	Interrogator's Features	Deceiver's Features
The Dec. Is Telling Truths	ACO	$MFCC_{2th}ofquartile1, linregc2, quartile2, amean$	$\Delta F0envofskewness, \Delta MFCC_7ofquartile3, Loudnessofstddev$
	CTD	$Ask_{ud}, Ask_{su}, Ask_{sd}$	-
	BERT	9 features	20 features
	POS	DM	-
The Dec. Is Telling Deceptions	ACO	$\Delta LSP_1ofskewness, VoiceProbofminPos$	$Loudnessofiqr1-3, \Delta ZCRamean$
	CTD	-	-
	BERT	20 features	10 features
	POS	DM	-

conversational temporal dynamics, and language use on truth-deception perception. We differ from [21], which only revealed that the prosodic and linguistic cues of deception and spotted some inconsistencies between the responses people perceived as deceptive and those that were actual deceptive statements.

We utilize acoustic-prosodic, conversational temporal dynamics, LIWC, part-of-speech tagging, and BERT representations that can be combined with EPQ scores to improve the automatic interrogator's perceived deception detection performances using a model based on BLSTM with attention mechanism network architecture. Our proposed framework achieves a promising accuracy of 86.70% and 84.89% (UAR) on 2-class *perceived* deception-truth recognition task on two conditions, the deceiver is telling the truths and deceptions, respectively. To the best of our knowledge, while there some research in studying speech perceived deception detection, this is one of the first studies that have explicitly modeled the personality traits together with acoustic-prosodic characteristics, linguistic cures, and language uses over the whole conversation on *perceived* deception and truth detection. Furthermore, we provide an analysis on the importance of different feature sets in perceived deception and truth detection on different conditions. In the immediate future work, we aim to spread out our multimodal fusion framework to combine multiple behaviors attributes to enhance the model robustness and the predicting powers by the early and late fusion, and to find the contribution made from an attention mechanism. Furthermore, we can directly model the four-class categories, which means we can know whether the interrogator is deceit successfully or not because we have both targets on each topic from the interrogator and the deceiver.

REFERENCES

[1] M. Hartwig, P. A. Granhag, L. A. Strmwall, and A. Vrij, "Police officers lie detection accuracy: Interrogating freely versus observing video," *Police Quarterly*, vol. 7, no. 4, pp. 429–456, 2004. [Online]. Available: https://doi.org/10.1177/1098611104264748

[2] A. Vrij, L. Akehurst, L. Brown, and S. Mann, "Detecting lies in young children, adolescents and adults," *Applied cognitive psychology*, vol. 20, no. 9, pp. 1225–1237, 2006. [Online]. Available: http://dx.doi.org/10.1002/acp.1278

[3] C. J. Dando and R. Bull, "Research article," *Journal of Investigative Psychology and Offender Profiling*, vol. 8, no. 2, pp. 189 – 202, 7 2011.

[4] A. L. Sandham, C. J. Dando, R. Bull, and T. C. Ormerod, "Improving professional observers veracity judgements by tactical interviewing," *Journal of Police and Criminal Psychology*, pp. 1–9, 2020.

[5] C. J. Dando, R. Bull, T. C. Ormerod, and A. L. Sandham, "Helping to sort the liars from the truth-tellers: The gradual revelation of information during investigative interviews," *Legal and Criminological Psychology*, vol. 20, no. 1, pp. 114–128, 2015. [Online]. Available: https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/lcrp.12016

[6] A. M. Crossman and M. Lewis, "Adults' ability to detect children's lying," *Behavioral sciences the law*, vol. 24, no. 5, p. 703715, 2006. [Online]. Available: https://doi.org/10.1002/bsl.731

[7] K. E. Sip, M. Lynge, M. Wallentin, W. B. McGregor, C. D. Frith, and A. Roepstorff, "The production and detection of deception in an interactive game," *Neuropsychologia*, vol. 48, no. 12, pp. 3619 – 3626, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0028393210003672

[8] G. Wright, C. Berry, and G. Bird, "You can't kid a kidder: association between production and detection of deception in an interactive deception task," *Frontiers in Human Neuroscience*, vol. 6, p. 87, 2012. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnhum.2012.00087

[9] K. Sip, D. Carmel, J. Marchant, J. Li, P. Petrovic, A. Roepstorff, W. McGregor, and C. Frith, "When pinocchio's nose does not grow: belief regarding lie-detectability modulates production of deception," *Frontiers in Human Neuroscience*, vol. 7, p. 16, 2013. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnhum.2013.00016

[10] D. Avola, L. Cinque, G. L. Foresti, and D. Pannone, "Automatic Deception Detection in RGB Videos Using Facial Action Units," in *Proceedings of the 13th International Conference on Distributed Smart Cameras*, ser. ICDSC 2019. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3349801.3349806

[11] I. Pavlidis and J. Levine, "Thermal Facial Screening for Deception Detection," in *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society][Engineering in Medicine and Biology*, vol. 2. IEEE, 2002, pp. 1143–1144.

[12] B. A. Rajoub and R. Zwiggelaar, "Thermal Facial Analysis for Deception Detection," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 6, pp. 1015–1023, 2014.

[13] R. S. Feldman, L. Jenkins, and O. Popoola, "Detection of Deception in Adults and Children via Facial Expressions," *Child Development*, vol. 50, no. 2, pp. 350–355, 1979. [Online]. Available: http://www.jstor.org/stable/1129409

[14] L. Su and M. D. Levine, "High-stakes Deception Detection Based on Facial Expressions," in *2014 22nd International Conference on Pattern Recognition*, 2014, pp. 2519–2524.

[15] W. Ambach and M. Gamer, "Chapter 1 - physiological measures in the detection of deception and concealed information," in *Detecting Concealed Information and Deception*, J. P. Rosenfeld, Ed. Academic Press, 2018, pp. 3 – 33. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B978012812729200001X

[16] F. A. Kozel, K. A. Johnson, E. L. Grenesko, S. J. Laken, S. Kose, X. Lu, D. Pollina, A. Ryan, and M. S. George, "Functional MRI detection of deception after committing a mock sabotage crime," *Journal of forensic sciences*, vol. 54, no. 1, p. 220231, January 2009. [Online]. Available: https://europepmc.org/articles/PMC2735094

[17] F. A. Kozel, S. J. Laken, K. A. Johnson, B. Boren, K. S. Mapes, P. S. Morgan, and M. S. George, "Replication of Functional MRI Detection of Deception," *Open forensic science journal*, vol. 2, no. 1, pp. 6–11, 2009. [Online]. Available: http://dx.doi.org/10.2174/1874402800902010006

[18] H. Chou, Y. Liu, and C. Lee, "JOINT LEARNING OF CONVERSATIONAL TEMPORAL DYNAMICS AND ACOUSTIC FEATURES FOR SPEECH DECEPTION DETECTION IN DIALOG GAMES," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 1044–1050.

[19] L. Zhou, D. P. Twitchell, Tiantian Qin, J. K. Burgoon, and J. F. Nunamaker, "An Exploratory Study into Deception Detection in Text-based Computer-mediated Communication," in *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the*, 2003, pp. 10 pp.–.

[20] A. Mbaziira and J. Jones, "A Text-based Deception Detection Model for Cybercrime," in *Int. Conf. Technol. Manag*, 2016.

[21] X. L. Chen, S. Levitan, M. Levine, M. Mandic, and J. Hirschberg, "Acoustic-Prosodic and Lexical Cues to Deception and Trust: Deciphering How People Detect lies," *Transactions of the Association for Computational Linguistics*, vol. 8, no. 0, pp. 199–214, 2020. [Online]. Available: https://transacl.org/ojs/index.php/tacl/article/view/1834

[22] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 14591462. [Online]. Available: https://doi.org/10.1145/1873951.1874246

[23] J. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. Booth, "The Development and Psychometric Properties of LIWC2007," 2011.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[25] P.-H. Li, T.-J. Fu, and W.-Y. Ma, "Why Attention? Analyze Bilstm Deficiency and Its Remedies in the Case of Ner," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[26] C.-H. Huang, H.-C. Chou, Y.-T. Wu, C.-C. Lee, and Y.-W. Liu, "Acoustic Indicators of Deception in Mandarin Daily Conversations Recorded from an Interactive Game," in *Proc. Interspeech 2019*, 2019, pp. 1731–1735. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2216

[27] S. D. Spencer, "Examining personality factors in deception detection ability," *Psi Chi Journal of Psychological Research*, vol. 22, no. 2, pp. 378–399, 2017.

[28] G. An, S. I. Levitan, J. Hirschberg, and R. Levitan, "Deep Personality Recognition for Deception Detection," in *Proc. Interspeech 2018*, 2018, pp. 421–425. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-2269

[29] C. Zhong-geng, "ITEM ANALYSIS OF EYSENCK PERSONALITY QUESTIONNAIRE TESTED IN BEIJING-DISTRICT," *Acta Psychologica Sinica*, vol. 2, 1983.

[30] V. Ivkovic, V. Vitart, I. Rudan, B. Janicijevic, N. Smolej-Narancic, T. Skaric-Juric, M. Barbalic, O. Polasek, I. Kolcic, Z. Biloglav, P. M. Visscher, C. Hayward, N. D. Hastie, N. Anderson, H. Campbell, A. F. Wright, P. Rudan, and I. J. Deary, "The Eysenck personality factors: Psychometric structure, reliability, heritability and phenotypic and genetic correlations with psychological distress in an isolated Croatian population," *Personality and Individual Differences*, vol. 42, no. 1, pp. 123 – 133, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0191886906002571

[31] H. J. Eysenck and S. B. G. Eysenck, *Manual of the Eysenck Personality Inventory: By HJ Eysenck and Sybil BG Eysenck*. University of London Press, 1964.

[32] H. J. Eysenck and S. G. B. Eysenck, "The eysenck personality inventory," *British Journal of Educational Studies*, vol. 14, no. 1, pp. 140–140, 1965.

[33] P. T. Barrett, K. V. Petrides, S. B. Eysenck, and H. J. Eysenck, "The eysenck personality questionnaire: An examination of the factorial similarity of p, e, n, and l across 34 countries," *Personality and Individual Differences*, vol. 25, no. 5, pp. 805–819, 1998.

[34] S. Mirsamadi, E. Barsoum, and C. Zhang, "AUTOMATIC SPEECH EMOTION RECOGNITION USING RECURRENT NEURAL NETWORKS WITH LOCAL ATTENTION," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2227–2231.

[35] S. I. Levitan, A. Maredia, and J. Hirschberg, "Acoustic-Prosodic Indicators of Deception and Trust in Interview Dialogues," in *Proc. Interspeech 2018*, 2018, pp. 416–420. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-2443

[36] tefan Beu, A. Gravano, and J. Hirschberg, "Pragmatic aspects of temporal accommodation in turn-taking," *Journal of Pragmatics*, vol. 43, no. 12, pp. 3001 – 3027, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378216611001469

[37] A. Vrij, S. Leal, L. Jupe, and A. 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, Sep. 2018.

[38] A. Vrij, M. Hartwig, and P. A. Granhag, "Reading Lies: Nonverbal Communication and Deception," *Annual Review of Psychology*, vol. 70, no. 1, pp. 295–317, 2019, pMID: 30609913. [Online]. Available: https://doi.org/10.1146/annurev-psych-010418-103135

[39] An Investigation on the Detectability of Deceptive Intent about Flying through Verbal Deception Detection, "An investigation on the detectability of deceptive intent about flying through verbal deception detection," *Collabra: Psychology*, vol. 3, no. 1, 2017.

[40] O. Litvinova, P. Seredin, T. Litvinova, and J. Lyell, "Deception Detection in Russian texts," in *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 43–52. [Online]. Available: https://www.aclweb.org/anthology/E17-4005

[41] S. I. Levitan, A. Maredia, and J. Hirschberg, "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)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1941–1950. [Online]. Available: https://www.aclweb.org/anthology/N18-1176

[42] E. Loper and S. Bird, "NLTK: The Natural Language Toolkit," in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ser. ETMTNLP '02. USA: Association for Computational Linguistics, 2002, p. 6370. [Online]. Available: https://doi.org/10.3115/1118108.1118117

[43] D. Klein and C. D. Manning, "Accurate Unlexicalized Parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ser. ACL '03. USA: Association for Computational Linguistics, 2003, p. 423430. [Online]. Available: https://doi.org/10.3115/1075096.1075150

[44] W.-Y. Ma and K.-J. Chen, "Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff," in *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing - Volume 17*, ser. SIGHAN '03. USA: Association for Computational Linguistics, 2003, p. 168171. [Online]. Available: https://doi.org/10.3115/1119250.1119276

[45] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception Detection Using Real-Life Trial Data," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 5966. [Online]. Available: https://doi.org/10.1145/2818346.2820758

[46] S. I. Levitan, G. An, M. Ma, R. Levitan, A. Rosenberg, and J. Hirschberg, "Combining Acoustic-Prosodic, Lexical, and Phonotactic Features for Automatic Deception Detection," in *Interspeech 2016*, 2016, pp. 2006–2010. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-1519

[47] M. Abouelenien, V. Pérez-Rosas, B. Zhao, R. Mihalcea, and M. Burzo, "Gender-Based Multimodal Deception Detection," in *Proceedings of the Symposium on Applied Computing*, ser. SAC '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 137144. [Online]. Available: https://doi.org/10.1145/3019612.3019644

[48] Y.-Y. Kao, P.-H. Chen, C.-C. Tzeng, Z.-Y. Chen, B. Shmueli, and L.-W. Ku, "Detecting Deceptive Language in Crime Interrogation," in *HCI in Business, Government and Organizations*, F. F.-H. Nah and K. Siau, Eds. Cham: Springer International Publishing, 2020, pp. 80–90.

[49] U. Avci and O. Aran, "Predicting the Performance in Decision-Making Tasks: From Individual Cues to Group Interaction," *IEEE Transactions on Multimedia*, vol. 18, no. 4, pp. 643–658, 2016.

[50] S.-C. Zhong, Y.-S. Lin, C.-M. Chang, Y.-C. Liu, and C.-C. Lee, "Predicting Group Performances Using a Personality Composite-Network Architecture During Collaborative Task," in *Proc. Interspeech 2019*, 2019, pp. 1676–1680. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2087

[51] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[52] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8026–8037. [Online]. Available: http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

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

[54] V. Hauch, I. Blandón-Gitlin, J. Masip, and S. L. Sporer, "Are computers effective lie detectors? A meta-analysis of linguistic cues to deception," *Personality and social psychology Review*, vol. 19, no. 4, pp. 307–342, 2015.