"Does it Matter When I Think You Are Lying?" Improving Deception Detection by Integrating Interlocutor's Judgements in Conversations

Huang-Cheng Chou[1,3], **Woan-Shiuan Chien**[1,3], **Da-Cheng Juan**[2], **Chi-Chun Lee**[1,3]
[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]Department of Computer Science, National Tsing Hua University, Taiwan
[3]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
{hc.chou,wschien,dacheng}@gapp.nthu.edu.tw, cclee@ee.nthu.edu.tw

Abstract

It is well known that human is not good at deception detection because of a natural inclination of truth-bias. However, during a conversation, when an interlocutor (interrogator) is being asked explicitly to assess whether his/her interacting partner (deceiver) is lying, this perceptual judgment depends highly on how the interrogator interprets the context of the conversation. While the deceptive behaviors can be difficult to model due to their heterogeneous manifestation, we hypothesize that this contextual information, i.e., whether the interlocutor trusts or distrusts what his/her partner is saying, provides an important condition in which the deceiver's deceptive behaviors are more consistently distinct. In this work, we propose a Judgmental-Enhanced Automatic Deception Detection Network (JEADDN) that explicitly considers interrogator's perceived truths-deceptions with three types of speech-language features (acoustic-prosodic, linguistic, and conversational temporal dynamics features) extracted during a conversation. We evaluate our framework on a large Mandarin Chinese Deception Dialog Database. The results show that the method significantly outperforms the current state-of-the-art approach without conditioning on the judgements of interrogators on this database. We further demonstrate that the behaviors of interrogators are important in detecting deception when the interrogators distrust the deceivers. Finally, with the late fusion of audio, text, and turn-taking dynamics (TTD) features, we obtain promising results of 87.27% and 94.18% accuracy under the conditions that the interrogators trust and distrust the deceivers in deception detection which improves 7.27% and 13.57% than the model without considering the judgements of interlocutor respectively.

1 Introduction

Deception behaviors frequently appear in human daily life, such as politics (Clementson, 2018), news (Conroy et al., 2015a; Vaccari and Chadwick, 2020), and business (Grazioli and Jarvenpaa, 2003; Triandis et al., 2001). Despite its frequent occurrences, researchers have repeatedly shown that humans are not good at detecting deceptions (it's 54% accuracy on average for both police officers and college students (Vrij and Graham, 1997)), even for highly-skilled professionals, such as teachers, social workers, and police officers (Hartwig et al., 2004; Vrij et al., 2006). Due to the difficulty in identifying deceptions by human, researchers have also developed an automatic deception detection (ADD) systems applied in various fields, such as cybercrime (Mbaziira and Jones, 2016), fake news detection (Conroy et al., 2015b), employment interviews (Levitan et al., 2018b,a), and even court decision (Venkatesh et al., 2019; Pérez-Rosas, Verónica and Abouelenien, Mohamed and Mihalcea, Rada and Burzo, Mihai, 2015). Although many works have studied approaches of automatic deception detection, few works, if any, has investigated whether judgements of human can help provide a condition that enhance ADD recognition rates.

In recent years, ADD has gained popularity and attention; however, almost all studies (if not all) on ADD pay attention to western cultures (countries), and there are very few literates that focus on eastern cultures (countries). Deception behavior often varies with different cultures (Aune and Waters, 1994), and every culture has its way to deceive others. Additionally, Rubin (2014) suggested that researchers need to study and understand more deception behaviors in the Asian area. Besides, many researchers have utilized various behavioral cues to build an ADD system, like facial expressions (Thannoon et al., 2019), internal physi-

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1846–1860

ological measures (Ambach and Gamer, 2018) and even functional brain MRI (Kozel et al., 2009a,b). While these indicators can be useful in detecting deceptions, many of them require expensive and invasive instrumentation that is not practical for real-world applications. Instead, speech and language cues carry substantial deceptive cues that can be modeled in ADD tasks for potential large-scale deployment (Zhou et al., 2003; Chou et al., 2019). Hence, the proposed method modeled the speech and language cues of humans with real-world data in Mandarin Chinese.

Despite these important advances in understanding and automatically identifying deceptions, there has been little work investigating whether the performance of ADD models can be significantly improved if considering the behaviors and perceptions of interrogators. Several questions remain: is there a difference in linguistic and acoustic-prosodic characteristics of an utterance from both interlocutors given trusted/distrusted judgments of interrogators? How do the judgments of interrogators help the ADD model detect deceptions? To investigate these questions, we firstly follow the previous studies (Chou et al., 2019) to segment a dialog into Questioning-Answering (QA) pair turns and then extract acoustic-prosodic features, linguistic features (e.g., Part-Of-Speech taggers (POS), Named Entity Recognition (NER), and Linguistic Inquiry and Word Count (LIWC)), conversational temporal dynamics (CTD) features. Then, we trained machine learning and deep learning classifiers using a large set of lexical and speech features to automatically identify deceptions and evaluated the results in the Daily Deceptive Dialogues corpus of Mandarin (DDDM). Also, to investigate the differences between interlocutor's behaviors, we perform Welch's t-test (Delacre et al., 2017) on the characteristics of utterances from both interlocutors given three different scenarios: (A) human-distrusted deceptive and truthful statements, (B) human-trusted deceptive and truthful statements, and (C) successful/unsuccessful deceptive and truthful statements.

In our further analyses, we found that (i) the judgments of human are indeed helpful to significantly improve the performance of the proposed method on detecting deceptions, (ii) the behaviors of interrogators should be considered into the model when the interrogator distrusted the deceivers, and (iii) the additional evidence indicates that human is bad at detecting deceptions – there are very few significant indicators that overlap between trusted truths-deceptions and successful-unsuccessful deceptions. We believe that these overlap-indicators could be useful for training humans to detect deceptions more successfully. Finally, we summarize our 3 main contributions as below.

- **We are the first work to include the judgements of the interrogator as a condition to help improve the recognition rates of deception detection model.**
- **We demonstrate that the features of interrogators are more effective and useful to detect deceptions than the deceivers' ones under the condition that the interrogator disbelieves the deceiver.**
- **The proposed model has high potentials for practical deception detection applications and impact on the ADD area.**

2 Related Work

Automatic deception detection in a dialogue Previous studies have trained a deception detector with various features in a dialog. Levitan et al. (2018a) extracted acoustic features of utterances to build the detection framework using a global-level label as the ground truth in employment interviews. Chou et al. (2019) indicated that the interlocutor's vocal characteristics and conversational dynamics should be jointly modeled to better perform deception detection in dialogues. The grammatical and syntactical POS features has been widely used in the automatic deception detection (Pérez-Rosas, Verónica and Abouelenien, Mohamed and Mihalcea, Rada and Burzo, Mihai, 2015; Levitan et al., 2016; Abouelenien et al., 2017; Kao et al., 2020). In addition, Liu et al. (2012); Levitan et al. (2018b) modeled the behaviors of language use from the LIWC features. Gröndahl and Asokan (2019); Chou et al. (2021) used textual embeddings extracted from the pre-trained BERT model for recognizing deceptions during an interrogator-deceiver conversation. Thannoon et al. (2019) used facial expression features to catch micro-variations on the face during the deceiver is telling either the lies or truths in the setting of interview conversation. Wu et al. (2018) had fused multimodal data including acoustic features, LIWC-embeddings, and facial-expression information to train a classifier for detecting deception, and Pérez-Rosas, Verónica and Abouelenien, Mohamed and Mihalcea, Rada and Burzo, Mihai (2015) trained the deception detec-

tion model using multimodal data with promising accuracy (92.20% area under the precision-recall curve, AUC) during a conversation in the court. However, most of the above-mentioned studies only model the behaviors of deceivers.

The interrogator's behaviors for detecting deceptions In criminal psychology, Dando and Bull (2011); Sandham et al. (2020) found that policies can be trained to identify criminal liars with advanced interrogation strategies (e.g, tactical use procedure) because these interview techniques maximize deceivers' cognitive load (Dando et al., 2015). In addition, Chou and Lee (2020) tried to learn from the behaviors of both interlocutors for identifying perceived deceptions, but their learning targets are from the perception of the interrogators not from the deceivers. Therefore, to our best knowledge, we are the first work to take the interrogators' behaviors for detecting deceptions automatically.

The perceptions of interrogators for detecting deceptions Levitan et al. (2018b) had studied the perception (judgment) of deception by identifying characteristics of statements that are perceived as truths or lies by interrogators, but they did not use the perceptions for detecting deceptions. Kleinberg and Verschuere (2021) used the LIWC variables and POS frequencies as input features to train a random forest classifier respectively, and then asked subjects to mark the scores ranging from 0 (certainty truthful) to 100 (certainty deceptive) on the deceptive or truthful text data. Finally, they presented the output probabilities of two trained classifiers on each data for the subjects to change the probabilities of the data. Their results showed that the perceptions of human impair the automatic deception detection models. However, we are different from Kleinberg and Verschuere (2021). The main difference is the way how judgements is being utilized; in our work, this is used to provide a condition in improving the prediction results.

Table 1: Distribution of the annotated data in the DDDM Database.

Data Distribution		Deceiver		Total
		Truth	Deception	
Interrogator	Trusted	(2) 97	(1) 86	**183**
	Distrusted	(3) 47	(4) 53	**100**

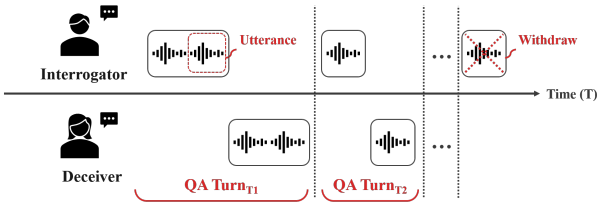

Figure 1: The illustration of Questioning-Answering (QA) pair turns. We only used *complete* QA pair turns and excluded some questioning turns if we cannot find the corresponding answering turns. To be noticed that each turn could have multiple utterances.

3 DDDM Database

We used conversational utterances from the Daily Deceptive Dialogues corpus of Mandarin (DDDM) (Huang et al., 2019). The entire DDDM contains about 27.2 hours of audio recordings from 96 unique speakers and 283 "question-level" conversational data samples. This corpus is particularly useful for our study, and all annotations in the DDDM come from "human" raters. Most deception databases lack recordings and perceptions (judgments) of the interrogators, while DDDM recorded the whole interrogator-deceiver conversations and the judgements of both interlocutors, allowing us to study deception detection given the judgements of the interrogators. With the judgements of both interlocutors, we group the data samples into four classes (shown in Table 1) as follows: (1) successful deceptions, (2) trusted truths, (3) distrusted truths, and (4) unsuccessful deceptions. We follow Chou et al. (2019) to transform the 7126 utterances into 2764 *complete* Questioning-Answering (QA) pair turns (shown in Figure 1) because the interrogator tended to ask follow-up questions for judging the deceiver's statements.

4 Problem Definition

4.1 The definition of deception

Deception is different from lying. Deception is human behavior that aims to make receivers believe true (or false) statements that the deceiver believes to be false (or true) with the conscious planning acts, such as sharing a mix of truthful and deceptive experiences to change the perceptions of the interrogators when being inquired to answer to questions. However, lying is just saying that something is true (or false) when in fact that something is false (or true) (Mitchell, 1986; Sarkadi, 2018). Hence, it is challenging for the interrogators to de-

tect deceptions through the behaviors of deceivers. Human needs to engage in higher-order cognitive processing to detect these consciously planned deceptions (Street et al., 2019). The deceiver can act in a way to change the perceptions of that potential deception detector. This then shifts a heavier burden onto the interrogator's cognitive processing. Hence, the interrogator must necessarily engage in "higher-order" cognitive processing to detect these advanced lies because they usually cannot just detect the behavior (e.g., signs of nervousness invoice), but must interpret why this individual may be nervous, including the honest reason why (e.g., afraid of being disbelieved).

4.2 Deception detection with judgments of human

Humans rarely perform better than chance on detecting deceptions, but the interrogators make their judgements according to context information in an interrogator-deceiver conversation. People might be hard to remember the whole detailed information, but their judgements might consist of some context-general information based on their own experience, which results in a truth-bias. Therefore, we build the deception detection models based on the conditional perceptions of humans (human-trusted or human-distrusted). We use judgements of human as criteria to define the following conditions (we also include the condition that we have no judgements of human, and the most conventional studies on ADD are in this condition):

(i) Truthful and deceptive statements detection: detecting deceptions without perceptions of interrogators (judgements of human)

(ii) Trusted truthful and deceptive statements detection: detecting deceptions with believed judgments of interrogators

(iii) Distrusted truthful and deceptive statements detection: detecting deceptions with disbelieved judgments of interrogators

5 Methodology

5.1 JEADDN: Judgmental-Enhanced Automatic Deception Detection Network

Figure 2 illustrates the Judgmental-Enhanced Automatic Deception Detection Network (JEADDN) whose main structure is BLSTM-DNN (Chou et al., 2019) containing one bidirectional long short-term memory (BLSTM) layer with an attention mechanism and two fully-connected layers. In our

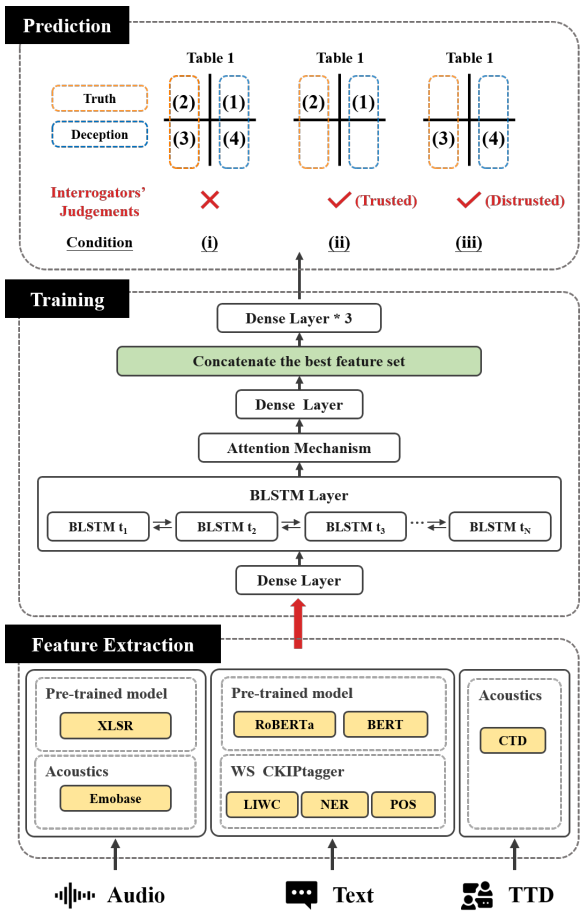

Figure 2: The overview of Judgmental-Enhanced Automatic Deception Detection Network (JEADDN) (The WS and TTD mean a word segmentation and turn-taking dynamics respectively).

method, judgements of human are criterion in choosing the classifiers for certain conditions to detect deceptions (not as the features). That is, when the interrogator believes the deceiver's statements, we use the condition (ii) classifier. Instead, when the interrogator disbelieves the deceiver, we can use the condition (iii) classifier. We fuse the best feature set from each modality by late fusion with additional three dense layers. Besides, there are two main goals. One is to investigate the effectiveness and robustness of speech and language features of both interlocutors. The other is to show whether the model performance of detecting deceptions with the judgements of interrogators could be better than the model without them.

More specifically, we split four-class sample data in Table 1 into two conditions based on judgements of interrogators (human-trusted/human-distrusted). The unit of features of interrogators/deceivers incorporates all of the utterances from the *complete*

Table 2: The table summarizes 8 turn-level feature sets introduction used in this paper.

Modality	Denotation	Feature Set	Dimension	Extraction Tool
Audio	Emobase	Emobase	988	openSmile (Eyben et al., 2010)
	XLSR	XLSR-53	7680	XLSR-53 (Conneau et al., 2020)
Text	NER	Named Entities Recognition	17	CKIP Tagger (Li et al., 2020)
	POS	Part-Of-Speech Tagger	50	CKIP Tagger (Li et al., 2020)
	BERT	BERT-Base in the Chinese version	768	BERT (Devlin et al., 2019)
	RoBERTa	RoBERTa-Base in the Chinese version	768	RoBERTa (Cui et al., 2020)
	LIWC	Linguistic Inquiry and Word Count 2015	82	LIWC 2015 (Pennebaker et al., 2015)
TTD	CTD	Conversational Temporal Dynamics	20	Proposed by Chou et al. (2019)

QA pair because interrogators would like to ask questions to seek detailed information. The closest previous study is Chou and Lee (2020). They have investigated **perceived** deception in the condition that the deceiver is telling either truths or deceptions, but they only focus on **perceived** deception recognition. Our objective is to detect the deceiver's answers corresponding to each question. In contrast, the learning targets of Chou and Lee (2020) are from the interrogator's guessed answers. Therefore, our learning targets are different from them. Moreover, their work is not useful in real life since they have to know the judgements of the deceivers, and it is impractical and impossible to be applied in the real world. In this paper, we hypothesize that (i) we can get better performance if the model takes judgements of interrogators into account, and (ii) there are differences in both interlocutors' behaviors between the trusted/distrusted truthful and deceptive dialogues. In the rest of the sections, we will describe the feature extraction in detail (notice that all types of the following feature sets are normalized to each speaker using z-score normalization) and the use of a deception detection framework.

5.1.1 Turn-level Feature Extraction

Table 2 summarized 8 various feature sets, which were extracted from the acoustic and linguistic characteristics of all speakers based on questioning turns of interrogators and answering turns of deceivers within QA pairs. In this work, we use the features extracted from audio and text recordings data to build the models, and we describe each feature set one by one as below.

Audio Recordings

- **Emobase:** we followed (Chou et al., 2019; Chou and Lee, 2020) to extracted 988-dimensional acoustic-prosodic features from questioning turns (answering turns) by "emobase.config" of openSMILE toolkit (Eyben et al., 2010).
- **CTD:** Chou et al. (2019) firstly proposed the conversational temporal dynamics (CTD) feature set within *complete* QA pairs. Additionally, Chou et al. (2021) incorporated CTD, Emobase, and BERT to achieve the state-of-the-art result of DDDM, so we also extract CTD for comparison. Also, CTD can extract the temporal turn-taking dynamics (TTD) of both interlocutors during a conversation. We list a part of features of CTD as below, and more detailed information about CTD is in Chou et al. (2019).
  - **Utterance-duration ratio:** the reciprocal ratio between the utterances length (u) and the turn duration (d), denoted as Int_{ud} and Int_{du} respectively.
  - **Silence-duration ratio:** the reciprocal ratio between the silence (s) duration and the turn duration, denoted as Int_{sd} and Int_{ds} respectively.
  - **Silence-utterance ratio:** the reciprocal ratio between the silence duration and the utterance lengths, denoted by Int_{su} and Int_{us} respectively.
  - **Silence times (st):** the number of times that a subject produces a pause that is more than 200ms, denoted as Int_{st} and Dec_{st}.
- **XLSR:** Due to the scarcity of deception databases in Mandarin Chinese, we use the multilingual pre-trained model, XLSR-53 (Conneau et al., 2020), to extract acoustic representation. XLSR-53 is trained for acoustic speech recognition (ASR) task with more than 56,000 hours of speech data in 53 different languages including Chinese-Taiwan (Mandarin Chinese) based on wav2vec 2.0 (Baevski et al., 2020). The dimension of the feature vector is 512 per frame, and then the feature vector per frame is applied

to the 15 statistics[1] to generate the final 7680-dimensional feature vectors.

Text Recordings

- **BERT:** we utilize BERT-Base in the Traditional Chinese version pre-trained model (Devlin et al., 2019) to extract turn-level 768-dimensional feature vectors. BERT was trained with a large amount of plain text data publicly available on the web using unsupervised objective functions (like masked-language modeling objective (MLM)) and works at the character level. We do not have to perform word segmentation when extracting representations.
- **RoBERTa:** we also use RoBERTa (Cui et al., 2020) to extract textual features. Its main component is the same as BERT (Devlin et al., 2019), but RoBERTa used a Whole-Word-Masking (WWM) technique and was trained on 10 times more data than BERT model. Although BERT has another version (BERT-WWM), there is no available pre-trained model in the Chinese language, so we only extract the features by both BERT and RoBERTa pre-trained models in this work.
- **POS:** we extracted 50-dimensional POS taggings excluding all punctuation-related dimensions by CKIP Tagger (Li et al., 2020) and then convert all POS predictions into feature vectors by calculating the number of word counts.
- **NER:** we use CKIP Tagger to extract 17-dimensional named entity recognition (NER) features, and convert the predictions into feature vectors by calculating the number of word counts. To our best knowledge, the NER feature set has never been used to train the deception detector. We are inspired by the findings of psychologist's studies on crime interrogation to use the NER feature set as input features for detecting deceptions. Vrij et al. (2021) suggest that the interrogators need to manipulate and design questions to ask the deceivers for detailed information, *complications*, because truth-tellers often reported more complications than lie tellers in each stage of the interview. A complication refers to details associated with personal experience or knowledge learned from any personal experience. In the DDDM, most recruited subjects are university students, and the three designed questions the researchers assigned each subject to ask are mainly about general activities or experiences of an average college student. For instance, scores of department border cups, professional knowledge about instruments, and detailed process of any events held by different clubs are regarded as personal experiences. Therefore, we extracted the NER features to capture the detailed information.
- **LIWC:** we use LIWC 2015 toolkit to extract 82-dimensional features (excluding all punctuation-related feature dimensions and total word counts (WP)) in this work after performing word segmentation pre-processing by CKIP Tagger.

6 Experiment

6.1 Experimental Setup

We conduct our experiments to show whether judgements and speech and language cues of interrogators are helpful to detect deceptions. The closest deception database is the Columbia X-Cultural Deception (CXD) Corpus (Levitan et al., 2015), but we have no access to the CXD corpus. To compare and show the baseline results, we compare all the models that had been used in the CXD corpus to reveal overall performance on the DDDM corpus. These baseline models include Support Vector Machines (SVM), Random Forest (RF), Logistic Regression (LG), and feedforward neural network (DNN). All baseline classifiers settings in this work are the same as Levitan et al. (2018b); Mendels et al. (2017).

Moreover, to compare with the state-of-the-art performance in the DDDM (Chou et al., 2021), we also use the same model proposed by (Chou et al., 2019), BLSTM-DNN, consisting of one fully-connected layer in a network with Rectified Linear Unit (ReLU) activation function, one BLSTM layer with an attention mechanism, one fully-connected layer with ReLU activation function, and then one prediction layer with a softmax activation function. We also include LSTM-DNN model in (Chou et al., 2021) as baseline classifier. All settings of LSTM-DNN and BLSTM-DNN are the same as (Chou et al., 2021). The whole framework is implemented by Pytorch (Paszke et al., 2019). The evaluation metric is macro F1-score based on the dyad-independent 10-fold cross-validation. We use the zero-padding to ensure each data sample's timestamp is the same if the length is less than the maximum timestamp (40). Several hyper-parameters for

[1](1): amean, (2): 1th percentile, (3): 99th percentile, (4): kurtosis, (5): 99th percentile minus first percentile, (6): max, (7): maxPos, (8): min, (9): minPos, (10): quartile1, (11): quartile3, (12): range, (13): skewness, (14): stddev, (15) median.

Table 3: Results on the produced deception detection on the DDDM database in macro F1-score (%). The **Who's Feature** column implies that the feature comes from whom, such as the interrogator (Int.), the deceiver (Dec.), or both of interlocutors (directly concatenate the features of interlocutors in feature-level).

Modality	**Feature**	**Who's Feature**	**(i)**						**(ii)**	**(iii)**
			RF (2018b)	LR (2018b)	SVM (2018b)	DNN (2018b)	LSTM-DNN (2019)	BLSTM-DNN (2019)	BLSTM-DNN	
Audio	Emobase	Int.	51.35	47.03	52.29	63.31	59.98	59.80	70.06	**83.72**
		Dec.	54.42	54.52	51.03	66.56	63.95	**66.84**	**76.92**	80.49
		Both	51.92	50.54	51.09	65.45	57.70	60.85	72.00	81.17
	XLSR	Int.	48.83	43.47	50.20	**65.02**	60.52	59.74	74.32	**83.10**
		Dec.	49.83	45.78	53.15	64.06	60.54	61.37	**75.25**	80.36
		Both	48.26	44.52	52.49	64.38	59.64	58.79	74.82	79.75
Text	NER	Int.	44.03	55.15	48.68	64.40	55.29	57.94	59.66	64.15
		Dec.	57.03	48.21	55.74	65.04	**68.10**	66.19	**74.78**	72.61
		Both	55.01	56.00	52.10	66.67	65.18	65.37	74.77	**75.61**
	POS	Int.	51.18	50.23	57.32	66.34	60.96	60.89	71.06	72.41
		Dec.	51.23	55.90	57.08	66.83	64.74	61.29	**75.19**	77.14
		Both	50.05	55.00	56.13	**67.09**	64.72	62.99	74.77	**77.32**
	LIWC	Int.	51.25	50.87	55.14	65.00	64.29	65.10	**76.27**	80.51
		Dec.	52.75	54.81	57.26	**68.36**	64.18	64.19	74.32	**82.44**
		Both	50.63	49.37	57.68	67.36	63.79	62.54	75.40	77.10
	BERT	Int.	54.61	58.61	54.47	68.30	65.98	63.15	77.38	**85.53**
		Dec.	60.77	62.38	57.76	71.03	70.83	**72.00**	**82.14**	82.77
		Both	61.99	62.62	57.69	71.05	70.82	71.63	77.52	83.30
	RoBERTa	Int.	52.43	53.00	55.93	69.04	66.96	65.78	73.46	80.80
		Dec.	56.22	59.22	57.79	70.45	73.13	**74.31**	**79.88**	**86.59**
		Both	58.21	61.33	61.10	71.35	73.17	73.56	75.43	81.83
TTD	CTD	Both	50.65	47.04	47.83	**65.11**	59.86	56.19	**71.90**	**64.00**

the LSTM-DNN and BLSTM-DNN models as below are grid-searched: the number of nodes in the LSTM and BLSTM layers is ranging in [2, 4, 8], and the batch size is ranging in [16, 32], the learning rates is ranging in [0.01, 0.005] with adjusting mechanism by multiplying $\frac{1}{\sqrt{1+epoch}}$ per epoch. Finally, the maximum epoch is 10000. These hyperparameters are chosen with early stopping criteria in all conditions to minimize cross-entropy with balanced class weights on the validation set.

6.2 Experimental Results

Table 3 presents a summary of the complete results in three different conditions. There are 283, 183, and 100 question-level data samples under conditions (i), (ii), and (iii) respectively. The more detailed information about the portion of DDDM is shown in Table 1. Besides, the human performance is 54.7% macro F1-score in the DDDM corpus. The performance of DNN (Mendels et al., 2017) is very competitive, but modeling time-series information is important for conversation setting. Hence, we only present the results with the BLSTM-DNN model in the conditions (ii) and (iii).

In Table 3, the performances of the BLSTM-DNN with judgments of interrogators are consistently higher than the models without the judgments of interrogators, and the findings show corroborating evidence of the ALIED theory (Street, 2015; Street et al., 2019) which claimed that the perceptions of human could be potential lie detector even though the judgments of human are error-prone. We also found that the interrogators' features seem more contributing to deception detection in condition (iii). This finding demonstrates that we could consider the interrogators' features when the interrogators distrust the deceivers for building deception detection models. However, the performances of most models trained with the feature sets of the deceivers in the condition (i) and (ii) consistently surpass the ones trained with the features from the interrogators or both interlocutors.

6.3 Ablation Study

To investigate the effectiveness of audio, text, and turn-taking dynamics (TTD) modalities, we take the feature set according to the best performance in Table 3. We take Emobase, BERT, and CTD to represent the audio, text, and TTD modalities respectively. In the condition (i) and (ii), Emobase and BERT are from the deceivers. On the other

Table 4: Ablation results on three modalities, Emobase, BERT, and CTD feature sets.

Modality	**Audio**	**Text**	**TTD**	**Condition**		
Feature Set	Emobase	BERT	CTD	(i)	(ii)	(iii)
Single Modality	V			66.84	76.92	83.72
		V		72.00	82.14	85.53
			V	56.19	71.90	64.00
Late Fusion	V	V		78.68	86.79	91.32
	V		V	74.92	84.99	87.63
		V	V	77.83	85.90	90.16
	V	V	V	80.61	87.27	94.18

hand, the counterparts are from the interrogators in the condition (iii). In the fusion method, we follow Chou et al. (2021) to firstly freeze the weights of all models trained with the above-mentioned feature sets and concatenate their final dense layers' outputs as the input of the additional three-layer feed-forward neural network to perform late fusion. Table 4 summarizes the results of the ablation study, and the text modality is the most effective modality. Finally, we get the promising results 87.27 % and 94.18 % and significant improvements 7.27% and 13.57% than the model without judgements of human in the condition (ii) and (iii) respectively.

7 Analyses

Having established the presence and characteristics of each speech and language cue, we were interested in exploring the differences in both of interlocutors' speech and language cues on the different judgements of the interrogators given three different scenarios: (A) human-distrusted deceptive and truthful statements, (B) human-trusted deceptive and truthful statements, and (C) successful/unsuccessful deceptive and truthful statements. We firstly performed Welch's t-test (Delacre et al., 2017) for each speaker's turn (e.g., questioning/answering turns) within QA pairs that represented a question and answer from the 3 daily questions. The QA pairs shown in Figure 1 were marked manually, and each deceivers' answer was labeled as truth or deception using the daily life questionnaire response sheet. This resulted in 2764 QA pairs. Using this data, the significant indicators after performing Welch's t-test between each feature set on the different conditions are shown in Appendix A.1 Table A.1. Then, we calculate the ratio of significant features in each feature set divided by its dimension base because every feature set has different dimensions, i.e., in the NER feature set under the scenario (A), there are 7 significant indicators and its dimension base is 17, so the ratio is calculated by 7 divided by 17. Additionally, while XLSR, NER, POS, BERT, and RoBERTa are all extracted by not zero-error-rate pre-trained models and LIWC is also calculated the word counts afterword segmentation by CKIP Tagger, they all have significant indicators whose p-value is smaller than 0.05 among them. For example, BERT and RoBERTa from the deceivers have a high proportion of significant indicators. However, since the meaning of XLSR, BERT, and RoBERTa representations are difficult to explain intuitively, so we focus on other feature sets to examine the following research questions.

Is there a difference in both interlocutors' behaviors between distrusted truths and deceptions (Scenario A)? According to the experimental results in Table 3, we understand that the features of interrogators are significant indicators to detect deceptions. After performing the Welch's t-test on each feature set between **distrusted** truthful and deceptive interlocutor's questioning/answering responses (there are 898 QA pairs in scenario A), we found that the feature sets of NER, POS, and LIWC have a higher ratio of statistically significant indicators. Moreover, we check the predictions of them in the DDDM, and we observe that the interrogators tend to ask more complex questions to inquire detailed information about the statements of deceivers. That is, the interrogators would check the numbers information about scores of games, frequency of presentation, or length of music concerts (*PERCENT*, *QUANTITY*, *Neqb*, and *DM*), things about musical instrument or events about concert presentations and ball games (*EVENT*, *PRODUCT*, and *WORK_OF_ART*), and places/locations (i.e., elementary schools and universities) (*Nc*). This result is very interesting because the psychologist studies had also shown that how interrogators interrogate the deceivers in details would affect the success in catching liars. Besides, there are some significant indicators in LIWC, such as the words describing the movements in the sport game (*death*: "殺"球 (殺球 means kill and spike)) and the words to ask the deceivers to provide more detailed information (*focusfuture*: "然後"你之後還有繼續打球/彈樂器嗎? ("then", did you keep playing balls/musical instrument afterward?)).

Is there a difference in both interlocutors' behaviors between trusted truths and deceptions (Scenario B)? In scenario B, the results of Welch's

t-test reveal that NER consists of the highest ratio of significant indicators than others. When we go back to read the data in the DDDM (Appendix A.1) Table A.1, we observe that the truthful statements have more detailed descriptions than the deceptive ones, such times/dates of ball games and concerts (*DATE* and *Nd*), numbers to describe the scores of games (*CARDINAL* and *PERCENT*), and names about musical instrument and sport equipment (*PRODUCT*). Besides, the significant indicators of Emobase shown in Appendix A.1 Table A.2 includes the first derivative of the intensity of the deceivers. This result is similar to the previous study on the English database (Chen et al., 2020). That is, the interrogator tended to judge high-intensity utterances as truths because the louder utterances might be perceived as more confident even though these utterances could be deceptive in fact. Additionally, the significance test shows that some CTD features of interrogators are important indicators indicating whether the deceiver is telling the truth or not when the interrogator trusted the deceivers. For example, in the Appendix A.2 Table A.3, we can find that the interrogator spends more time to come up with more complex questions to inquire the deceiver; however, the interrogator eventually believes the deceiver's statements, but the proposed method can successfully detect the deceptions by the interrogator's temporal TTD behaviors. This finding is the same as the previous study (Chou et al., 2019).

Is there any common significant indicator between the one from distrusted truths and deceptions and the other from successful/unsuccessful deceptions (Scenario C)? In this analysis, we demonstrate additional evidence indicating that human is poor at detecting deceptions–there are very few indicators that overlap in all feature sets in this condition in Appendix A.1 Table A.1 (the rightmost column). However, the results repeatedly show that the ways how the interrogators ask questions about detailed information (*MONEY*, *PRODUCT*, and *DM*), and the meaningful information in the deceivers' answering statements (*A* (one of POS features) means the words to describe the noun, such as female, big, small, to name a few). Hence, the more detailed information we have, the higher chances to detect deceptions.

8 Conclusion and Future Work

This paper investigates whether judgements and speech and language cues of interrogators in conversation are useful and helpful to detect deceptions. We analyzed a full suite of acoustic-prosodic features, linguistic cues, conversational temporal dynamics given different conditions. Finally, with the late fusion of audio, text, and turn-taking dynamics (TTD) modality features, JEADDN obtains promising results of 87.27% and 94.18% accuracy under the conditions that the interrogators trust and distrust the deceivers in deception detection which improves 7.27% and 13.57% than the model without considering the interlocutor's judgements respectively.

While there is some research in studying perceived deception detection, this is one of the first studies that have explicitly modeled acoustic-prosodic characteristics, linguistic cues, and conversational temporal dynamics using judgments of interrogators in conversations for detecting deceptions. Furthermore, we provide analyses on the significance of different feature sets in three different scenarios and show additional evidence indicates that human is bad at detecting deceptions. Especially, the content of questions the interrogators ask is an indicator for telling deceptions or truths when the interrogators distrust the deceivers. Verigin et al. (2019) also reveal that truthful and deceptive information interacts to influence detail richness provides insight into liars' strategic manipulation of information when statements contain a mixture of truths and lies.

In the immediate future work, we aim to extend our multimodal fusion framework to combine semantic information to enhance the model robustness and the predicting powers within multiple QA pairs. That is, we observe that some interrogators finally trusted the deceivers after many follow-up questions while the statements of the deceivers were deceptive. Kontogianni et al. (2020) also pointed out that follow-up open-ended questions prompt additional reporting. However, practitioners should be cautious to corroborate the accuracy of new reported details.

Acknowledgments

The work is supported in part by funding from the Ministry of Science and Technology, Taiwan. We thank the anonymous reviewers, Jason S. Chang, and Jhih-Jie Chen for their valuable comments.

References

Mohamed Abouelenien, Verónica Pérez-Rosas, Bohan Zhao, Rada Mihalcea, and Mihai Burzo. 2017. Gender-Based Multimodal Deception Detection. In *Proceedings of the Symposium on Applied Computing*, SAC '17, page 137–144, New York, NY, USA. Association for Computing Machinery.

Wolfgang Ambach and Matthias Gamer. 2018. Chapter 1 - physiological measures in the detection of deception and concealed information. In J. Peter Rosenfeld, editor, *Detecting Concealed Information and Deception*, pages 3 – 33. Academic Press.

R.Kelly Aune and Linda L. Waters. 1994. Cultural differences in deception: Motivations to deceive in Samoans and North Americans. *International Journal of Intercultural Relations*, 18(2):159 – 172.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12435–12446. Curran Associates, Inc.

Xi (Leslie) Chen, Sarah Ita Levitan, Michelle Levine, Marko Mandic, and Julia Hirschberg. 2020. Acoustic-Prosodic and Lexical Cues to Deception and Trust: Deciphering How People Detect Lies. *Transactions of the Association for Computational Linguistics*, 8:199–214.

Huang-Cheng Chou and Chi-Chun Lee. 2020. "Your Behavior Makes Me Think It Is a Lie": Recognizing Perceived Deception using Multimodal Data in Dialog Games. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 393–402.

Huang-Cheng Chou, Yi-Wen Liu, and Chi-Chun Lee. 2019. 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)*, pages 1044–1050.

Huang-Cheng Chou, Yi-Wen Liu, and Chi-Chun Lee. 2021. Automatic deception detection using multiple speech and language communicative descriptors in dialogs. *APSIPA Transactions on Signal and Information Processing*, 10:e5.

David E. Clementson. 2018. Truth Bias and Partisan Bias in Political Deception Detection. *Journal of Language and Social Psychology*, 37(4):407–430.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised Cross-lingual Representation Learning for Speech Recognition.

Nadia K. Conroy, Victoria L. Rubin, and Yimin Chen. 2015a. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Nadia K. Conroy, Victoria L. Rubin, and Yimin Chen. 2015b. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Coral J. Dando and Ray Bull. 2011. Research article. *Journal of Investigative Psychology and Offender Profiling*, 8(2):189 – 202.

Coral J. Dando, Ray Bull, Thomas C. Ormerod, and Alexandra L. Sandham. 2015. Helping to sort the liars from the truth-tellers: The gradual revelation of information during investigative interviews. *Legal and Criminological Psychology*, 20(1):114–128.

M. Delacre, D. Lakens, and C. Leys. 2017. Why psychologists should by default use welch's t-test instead of student's t-test. *International Review of Social Psychology*, 30(1):92–101.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. MM '10, New York, NY, USA. Association for Computing Machinery.

Stefano Grazioli and Sirkka L. Jarvenpaa. 2003. Consumer and Business Deception on the Internet: Content Analysis of Documentary Evidence. *Int. J. Electron. Commerce*, 7(4):93–118.

Tommi Gröndahl and N. Asokan. 2019. Text Analysis in Adversarial Settings: Does Deception Leave a Stylistic Trace? 52(3).

Maria Hartwig, Pär Anders Granhag, Leif A. Strömwall, and Aldert Vrij. 2004. Police Officers' Lie Detection Accuracy: Interrogating Freely Versus Observing Video. *Police Quarterly*, 7(4):429–456.

Chih-Hsiang Huang, Huang-Cheng Chou, Yi-Tong Wu, Chi-Chun Lee, and Yi-Wen Liu. 2019. Acoustic Indicators of Deception in Mandarin Daily Conversations Recorded from an Interactive Game. In *Proc. Interspeech 2019*, pages 1731–1735.

Yi-Ying Kao, Po-Han Chen, Chun-Chiao Tzeng, Zi-Yuan Chen, Boaz Shmueli, and Lun-Wei Ku. 2020. Detecting Deceptive Language in Crime Interrogation. In *HCI in Business, Government and Organizations*, pages 80–90, Cham. Springer International Publishing.

Bennett Kleinberg and Bruno Verschuere. 2021. How humans impair automated deception detection performance. *Acta Psychologica*, 213:103250.

Feni Kontogianni, Lorraine Hope, Paul J. Taylor, Aldert Vrij, and Fiona Gabbert. 2020. "tell me more about this…": An examination of the efficacy of follow-up open questions following an initial account. *Applied Cognitive Psychology*, 34(5):972–983.

F Andrew Kozel, Kevin A Johnson, Emily L Grenesko, Steven J Laken, Samet Kose, Xinghua Lu, Dean Pollina, Andrew Ryan, and Mark S George. 2009a. Functional MRI detection of deception after committing a mock sabotage crime. *Journal of forensic sciences*, 54(1):220—231.

F. Andrew Kozel, Steven J. Laken, Kevin A. Johnson, Bryant Boren, Kimberly S. Mapes, Paul S. Morgan, and Mark S. George. 2009b. Replication of Functional MRI Detection of Deception. *Open forensic science journal*, 2(1):6–11.

Sarah I. Levitan, Guzhen An, Mandi Wang, Gideon Mendels, Julia Hirschberg, Michelle Levine, and Andrew Rosenberg. 2015. Cross-Cultural Production and Detection of Deception from Speech. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, WMDD '15, page 1–8, New York, NY, USA. Association for Computing Machinery.

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 2016*, pages 2006–2010.

Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg. 2018a. Acoustic-Prosodic Indicators of Deception and Trust in Interview Dialogues. In *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)*, pages 1941–1950, New Orleans, Louisiana. Association for Computational Linguistics.

Peng-Hsuan Li, Tsu-Jui Fu, and Wei-Yun Ma. 2020. Why Attention? Analyze Bilstm Deficiency and Its Remedies in the Case of Ner. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

X. Liu, K. Tang, J. Hancock, J. Han, M. Song, R. Xu, V. Manikonda, and B. Pokorny. 2012. Socialcube: A text cube framework for analyzing social media data. In *International Conference on Social Informatics (SocialInformatics)*, pages 252–259, Los Alamitos, CA, USA. IEEE Computer Society.

A Mbaziira and J Jones. 2016. A Text-based Deception Detection Model for Cybercrime. In *Int. Conf. Technol. Manag*.

Gideon Mendels, Sarah Ita Levitan, Kai-Zhan Lee, and Julia Hirschberg. 2017. Hybrid Acoustic-Lexical Deep Learning Approach for Deception Detection. In *Proc. Interspeech 2017*, pages 1472–1476.

Robert W Mitchell. 1986. A framework for discussing deception. *Deception: Perspectives on human and nonhuman deceit*, pages 3–40.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, Inc.

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

Pérez-Rosas, Verónica and Abouelenien, Mohamed and Mihalcea, Rada and Burzo, Mihai. 2015. Deception Detection Using Real-Life Trial Data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, page 59–66, New York, NY, USA. Association for Computing Machinery.

Victoria L. Rubin. 2014. TALIP Perspectives, Guest Editorial Commentary: Pragmatic and Cultural Considerations for Deception Detection in Asian Languages. *ACM Transactions on Asian Language Information Processing*, 13(2).

Alexandra L Sandham, Coral J Dando, Ray Bull, and Thomas C Ormerod. 2020. Improving Professional Observers' Veracity Judgements by Tactical Interviewing. *Journal of Police and Criminal Psychology*, pages 1–9.

Stefan Sarkadi. 2018. Deception. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 5781–5782. AAAI Press.

Chris N. H. Street, Jaume Masip, and Megan Kenny. 2019. *Understanding Lie Detection Biases with the*

Adaptive Lie Detector Theory (ALIED): A Boundedly Rational Approach, pages 227–247. Springer International Publishing, Cham.

Chris NH Street. 2015. Alied: Humans as adaptive lie detectors. *Journal of Applied Research in Memory and Cognition*, 4(4):335 – 343.

Harith H Thannoon, Wissam H Ali, and Ivan A Hashim. 2019. Design and Implementation of Deception Detection System Based on Reliable Facial Expression. *Journal of Engineering and Applied Sciences*, 14(15):5002–5011.

Harry C. Triandis, Peter Carnevale, Michele Gelfand, Christopher Robert, S. Arzu Wasti, Tahira Probst, Emiko S. Kashima, Thalia Dragonas, Darius Chan, Xiao Ping Chen, Uichol Kim, Carsten De Dreu, Evert Van De Vliert, Sumiko Iwao, Ken-Ichi Ohbuchi, and Paul Schmitz. 2001. Culture and Deception in Business Negotiations: A Multilevel Analysis. *International Journal of Cross Cultural Management*, 1(1):73–90.

Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, 6(1):2056305120903408.

Sushma Venkatesh, Raghavendra Ramachandra, and Patrick Bours. 2019. Robust Algorithm for Multimodal Deception Detection. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 534–537.

Brianna L. Verigin, Ewout H. Meijer, Aldert Vrij, and Leonie Zauzig. 2019. The interaction of truthful and deceptive information. *Psychology, Crime & Law*.

Aldert Vrij, Lucy Akehurst, Laura Brown, and Samantha Mann. 2006. Detecting lies in young children, adolescents and adults. *Applied cognitive psychology*, 20(9):1225–1237.

Aldert Vrij, Samantha Mann, Sharon Leal, and Ronald P. Fisher. 2021. Combining verbal veracity assessment techniques to distinguish truth tellers from lie tellers. *The European Journal of Psychology Applied to Legal Context*, 13(1):9–19.

Dr Aldert Vrij and Sally Graham. 1997. Individual differences between liars and the ability to detect lies. *Expert Evidence*, 5(4):144–148.

Zhe Wu, Bharat Singh, Larry Davis, and V. Subrahmanian. 2018. Deception Detection in Videos.

Lina Zhou, Douglas P. Twitchell, Tiantian Qin, Judee K. Burgoon, and Jay F. Nunamaker. 2003. 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*, pages 10 pp.–.

A Appendix

A.1 Welch's T-test Results

The Welch's t-test results are shown in Table A.1 and Table A.2 based on speakers' turns within QA pairs in three different scenarios as follows: (A) human-distrusted deceptive and truthful statements, (B) human-trusted deceptive and truthful statements, and (C) successful/unsuccessful deceptive and truthful statements. Emobase contains the fundamental frequency (F0) and its envelope, intensity (*INTENSITY*), loudness (*LOUDNESS*), 12 MFCC, probability of voicing (*VOICEPROB*), 8 line spectral frequencies (*LSPFREQ*), zero-crossing rate (*ZCR*), and delta regression coefficients. Then, these LLDs and their delta coefficients are applied to the following statistics [2] to generate the final feature vector. The number (ratio) represents the number of signatures in each feature set divided by its dimension base because every feature set has different dimensions. For instance, in the NER feature set under scenario (A), there are 7 significant indicators and its dimension base is 17, so the ratio (number) is calculated by 7 divided by 17.

A.2 A Real Example in the DDDM database

Table A.3 summarizes a real example in the DDDM database, and we also show its duration, transcripts in Mandarin Chinese, and translation in English. This data sample is grouped into (1) class in Table 1. The interrogator trusts the deceiver but the deceiver tells deception in fact.

[2](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): skewness, (19): stddev.

Table A.1: The results of all feature sets after performing Welch's t-test in three different scenarios: (A) human-distrusted deceptive and truthful statements, (B) human-trusted deceptive and truthful statements, and (C) successful/unsuccessful deceptive and truthful statements ("*" indicates the significance threshold, p-value, is smaller than 0.01; "**" is smaller than 0.001).

Modality	Feature	Who's Feature	(A) (%)	Indicators	(B) (%)	Indicators	(C) (%)	Indicators	(B)⋂(C) (%)	Indicators
Audio	Emobase	Int.	4.37	Table A.2	1.02	Table A.2	1.83	Table A.2	0.10	$\Delta\text{MFCC}^{4\text{th}}$ – skewness
		Dec.	1.93		4.37		4.68		0.10	$\Delta\text{MFCC}^{7\text{th}}$ – linregc2
	XLSR	Int.	3.07	-	3.75	-	2.79	-	0.44	-
		Dec.	2.21	-	3.54	-	2.92	-	0.52	-
Text	NER	Int.	**29.41**	EVENT**, PRODUCT**, WORK_OF_ART**, PERCENT, QUANTITY	11.76	MONEY*, PRODUCT*	**35.29**	PRODUCT**, EVENT*, MONEY*, PERCENT, QUANTITY, WORK_OF_ART	**11.76**	MONEY*, PRODUCT*
		Dec.	5.88	PERCENT	23.53	PERCENT**, DATE*, CARDINAL, PRODUCT	11.76	PERCENT**, WORK_OF_ART**	5.88	PERCENT
	POS	Int.	14.00	Neqb*, D, DM, Dfb, Nc, VCL, VA	10.00	DM**, VG**, Da*, Nb, A	6.00	DM**, Dfb**, VI**	2.00	DM
		Dec.	0.00	-	10.00	DM**, A, FW, Nd, V_2	8.00	A, Cbb, Dk, V_2	4.00	A, V_2
	LIWC	Int.	10.98	death**, adverb* leisure, cogproc, focusfuture*, filler, auxverb, discrep, othergram	2.44	you, cogproc	7.32	death**, negate, filler, ipron, I, Sixltr	0.00	-
		Dec.	1.22	female	3.66	bio**, sexual, power	3.66	you, informal, social	0.00	-
	BERT	Int.	6.51	-	4.43	-	8.46	-	0.26	-
		Dec.	10.55	-	18.36	-	3.52	-	0.78	-
	RoBERTa	Int.	8.59	-	5.99	-	9.51	-	0.65	-
		Dec.	10.03	-	17.45	-	6.77	-	1.30	-
TTD	CTD	Both	0.00	-	**30.00**	Int_{sd}*, Int_{ud}*, Int_{su}*, Int_{us}, $\text{Int}_{\text{d}}/\text{Dec}_{\text{d}}$, Int_{st}	15.00	Dec_{ud}, Dec_{sd}, Dec_{us}	0.00	-

Table A.2: The Welch's t-test results on Emobase in three different scenarios ("*" indicates the significance threshold, p-value, is smaller than 0.01; "**" is smaller than 0.001).

Scenario	Interrogator	Deceiver
(C)	MFCC^{3th}: (6*), MFCC^{8th}: (12, 18), MFCC^{9th}: (8, 9, 12, 19), MFCC^{10th}: (1, 10, 14, 16), ΔMFCC^{4th}: (18*), ΔMFCC^{5th}: (1), ΔMFCC^{11th}: (14), LSPFREQ^{1th}: (1, 12), ΔLSPFREQ^{1th}: (15*, 18), LOUDNESS: (1, 12)	MFCC^{1th}: (4*, 15), MFCC^{3th}: (3, 4, 8, 9, 19), MFCC^{5th}: (14), MFCC^{8th}: (1, 16), MFCC^{9th}: (1, 15, 16), MFCC^{11th}: (6, 7), ΔMFCC^{1th}: (7), ΔMFCC^{7th}: (7), ΔMFCC^{11th}: (1. 14), LSPFREQ^{0th}: (2, 3*, 4, 16), LSPFREQ^{1th}: (1, 7, 15), LSPFREQ^{2th}: (1, 14, 15, 16), LSPFREQ^{6th}: (7), LSPFREQ^{7th}: (2*, 3, 8, 9*, 19*), ΔLSPFREQ^{7th}: (18)VOICEPROB: (16), ZCR: (2*, 15), F0: (12*), ΔF0: (3, 4, 8*, 9, 14, 19)
(B)	MFCC^{4th}: (10), MFCC^{6th}: (18), MFCC^{7th}: (9), ΔMFCC^{4th}: (18), ΔMFCC^{8th}: (6, 7*), ΔVOICEPROB (4), ΔLSPFREQ^{4th}: (18), ΔLSPFREQ^{7th}: (6, 7)	MFCC^{2th}: (18), MFCC^{6th}: (8, 9, 12, 17, 18, 19*), MFCC^{8th}: (2*, 3, 8*, 9*, 12*, 17, 19*), MFCC^{9th}: (2*, 3, 8, 9, 18, 19*), MFCC^{10th}: (5, 18), ΔMFCC^{6th}: (9, 12, 19), ΔMFCC^{7th}: (6*, 7*), ΔMFCC^{8th}: (1, 6, 7, 8*, 9*, 10*, 12, 16, 17*, 19*), ΔMFCC^{12th}: (10), LSPFREQ^{3th}: (5, 18), ΔLSPFREQ^{7th}: (7, 15), ΔINTENSITY: (18)
(A)	MFCC^{1th}: (12*, 17*), MFCC^{2th}: (1*, 7, 12*, 14*, 15*, 16, 17), MFCC^{3th}: (1, 6), MFCC^{8th}: (12), MFCC^{9th}: (10, 18), MFCC^{10th}: (1, 2, 3, 14), MFCC^{11th}: (12), MFCC^{12th}: (6), ΔMFCC^{1th}: (12, 18), ΔMFCC^{3th}: (15), ΔMFCC^{4th}: (9), ΔMFCC^{6th}: (7), ΔMFCC^{9th}: (13, 14), ΔMFCC^{10th}: (14), LSPFREQ^{0th}: (1), LSPFREQ^{1th}: (10), LSPFREQ^{3th}: (10), LSPFREQ^{4th}: (12, 17*), LSPFREQ^{5th}: (12, 17), LSPFREQ^{7th}: (5, 18), ΔLSPFREQ^{1th}: (18*), ΔLSPFREQ^{5th}: (12), VOICEPROB: (12, 19), ΔVOICEPROB: (10, 17)	MFCC^{2th}: (14), MFCC^{3th}: (2), MFCC^{4th}: (1, 14), MFCC^{5th}: (7), MFCC^{6th}: (9, 10, 17, 19*), MFCC^{8th}: (14), MFCC^{9th}: (18), MFCC^{11th}: (6), ΔMFCC^{6th}: (1, 2, 9), LSPFREQ^{0th}: (3), ΔLSPFREQ^{7th}: (2, 14), ΔF0: (18)

Table A.3: A Real Example in the DDDM database.

(Questioning Turn) (Duration) Interrogator's Questing Turns	**(Answering Turn) (Duration) Deceiver's Answering Turns**
(Q1)(09.6s)那因為你寫高中後有參加過吉他的公開比賽 或演奏那請問是怎樣類型的公開 (According to your answer sheet, you have participated in public guitar competitions or performances after high school. What kind of public is that?)	(A1)(10.368s)其實就是呃一般類就是社團的那個發表會 這樣主要是我那個大學的時候 (In fact, er, it's a general club's presentation, mainly when I was in university.)
(Q2)(01.92s)嗯大概是大幾的時候 (Um, what was your grade at tha time?)	(A2)(00.768s)大一的時候 (Freshman.)
(Q3)(04.864s)那那時候你是吉他那那你是吉他社嗎 (At that time, were you a member of a guitar club?)	(A3)(14.336s)大一的時候算吧因為就是大一的時候 就是比較比較時間比較多所以去了蠻多社團所以 也有去吉他社就是呃自己重頭開始練然後有參加過 一個學期的就是成發這樣 (When I was a freshman, I have more time in the freshman year, so I went to a lot of clubs, so I also went to the guitar club, um, I started practicing on my own and then participated in presentation for a semester.)
(Q4)(08.704s)那那表演的時候是你個人獨秀 還是大家一起彈的還是樂團 (So when you perform, is it your solo show , or is it the orchestra that everyone plays together?)	(A4)(12.928)主要是我跟另外一個就是社團的朋友就 我們兩個呃我就是他彈我彈欸我彈主旋律他彈就是節奏 這樣然後vocal的話就是一起 (Another friend and I from the club. Uh, he plays, I play, I play the main melody, he plays the rhythm, and the vocal is together.)
(Q5)(06.272s)那那目前你還有在繼續練吉他嗎 (Are you still practicing guitar?)	(A5)(08.064)後來就沒有了後來就是就比較喜歡去熱舞社 這樣所以吉他社就沒有再去了 (Later, no. Later, I prefer to go to the hot dance club, so the guitar club I did not go.)
(Q6)(13.184s) 嗯那能簡單講一下吉他的基本入門五大 和弦嗎就是最常出現的那幾個 (Um, can you briefly talk about the basic and common guitar five major chords that appear most often?)	(A6)(08.704)有點忘了我知道有C然後C1G吧 我只記得這幾個對 (I forgot. I know that there are C and then C1G. I only remember these pairs. Yes.)