SIP (2021), vol. 10, e5, page 1 of 9 © The Author(s), 2021 published by Cambridge University Press in association with Asia Pacific Signal and Information Processing Association.. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.
doi:10.1017/ATSIP.2021.6

ORIGINAL PAPER

Automatic Deception Detection using Multiple Speech and Language Communicative Descriptors in Dialogs

HUANG-CHENG CHOU,[1,2] YI-WEN LIU[1] AND CHI-CHUN LEE[1,2]

While deceptive behaviors are a natural part of human life, it is well known that human is generally bad at detecting deception. In this study, we present an automatic deception detection framework by comprehensively integrating prior domain knowledge in deceptive behavior understanding. Specifically, we compute acoustics, textual information, implicatures with non-verbal behaviors, and conversational temporal dynamics for improving automatic deception detection in dialogs. The proposed model reaches start-of-the-art performance on the Daily Deceptive Dialogues corpus of Mandarin (DDDM) database, 80.61% unweighted accuracy recall in deception recognition. In the further analyses, we reveal that (i) the deceivers' deception behaviors can be observed from the interrogators' behaviors in the conversational temporal dynamics features and (ii) some of the acoustic features (e.g. loudness and MFCC) and textual features are significant and effective indicators to detect deception behaviors.

Keywords: Automatic deception detection, Implicature classification, Conversational temporal dynamics, BLSTM with hierarchical attention networks

Received 23 November 2020; Revised 17 March 2021

I. INTRODUCTION

Deception is a planned intentional behavior of a deceiver to make an interrogator believe a statement to be true or false when the deceiver already knows it to be false or true, respectively. Deception is often mechanistically used to share a mix of truthful and deceptive experiences when being inquired and interrogated [1]. Although deception behaviors frequently exist in our daily life, such as in politics [2], news [3, 4], and business settings [5, 6], it is challenging for untrained personnel to identify deception accurately. According to [7], deception detection accuracy is only at 54% on average for both police officers and college students. Further, there is a known phenomenon termed "truth-bias", i.e. people often turn to believe strangers' statements [8, 9]. Hence, the ability to consistently detect deception with high reliability is important in many application fields, e.g. fake news detection [3], employment interviews [10, 11], and even court decisions [12, 13].

Moreover, while deception behaviors often vary with different cultures [14], i.e. each culture has its way of expressing deception, most of the studies on automatic deception detection focus on western cultures (countries). Very few studies have investigated methods in developing deception detection for eastern cultures (countries). A recent study done by Rubin [15] suggests that researchers should pay closer attention to the deception behavior in the Asian culture as it may be drastically different from the much well-studied western culture. In this work, we aim to investigate the methods of automatic deception detection for Mandarin Chinese native speakers.

One of the most common deceptions occurring in situations is in conversation settings. Many researchers have computed a variety of behavioral cues to build automatic deception detection during conversations. For example, Levitan *et al.* [11] extracted utterance-like low-level descriptors (e.g. acoustic features) to train the detection framework for settings of employment interviews. Thannoon *et al.* [16] used facial expression features to characterize micro-variations of the deceiver's face during interview conversation. Other literature also utilized behaviors of language use (e.g. features derived from Linguistic Inquiry and Word Count (LIWC) [10, 17] or pre-trained BERT model [18]) to train deception classifiers to be used in a conversation setting. Lastly, several types of research have investigated fusion methods of multimodal behavior data, including acoustic features, LIWC-embeddings, and facial expressions, for automatic deception detection [19]. By utilizing multimodal data, the detection model [13] was shown to be

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

Corresponding author:
Huang-Cheng Chou
Email: hc.chou@gapp.nthu.edu.tw
Chi-Chun Lee
Email: cclee@ee.nthu.edu.tw

capable of obtaining high accuracy (92.20% Area under the Curve of ROC, AUC) for conversations in the court.

Yet, there are a wide range of important aspects to be considered in conversation: acoustic-prosodic patterns, turn-taking dynamics, lexical semantics, and higher-level information of an utterance in a dialog (e.g. pragmatics and implicatures). Most prior studies only focus on acoustic-prosodic variations, LIWC, or BERT text-embeddings to be the input to the detection models. While many behavior science studies, e.g. those in psychology, social science, and conversational linguistics, have already shown many of these aspects are related to the expression of deceptions, very few computational works, if any, have explicitly considered these to improve the automatic deception model. Only recently, Chou et al. [20] showed that conversation temporal dynamics can be integrated as input features to help improve deception detection, where many of their derived features are inspired by studies of conversation scholars [21]. In this work, our goal is to integrate comprehensive speech and language communicative descriptors into constructing an automatic deception detection model. Numerous studies have identified the relationship between deception and the variability in the temporal alignment of turn initiations and the pragmatics of interpersonal communication (e.g. backchannel, unnormal pauses, stammer). Furthermore, the implicature in the speaking content of the deceiver can be categorized into three classes including complication, common knowledge details, and self-handicapping; studies have shown that the use of these implicatures is useful to improve a human's ability to detect deceptions [22, 23, 23–26]. We would explicitly consider these two broad categories of speech/language attributes in dialogs along with the conventional acoustic-prosodic and textual information to perform deception detection.

The previous work [20] was the first in modeling the conversational temporal dynamics based on dyadic turn-taking behaviors. To be more specific, they designed 20-dimensional temporal features in each "questioning–answering" (QA) pair composed of a questioning turn and an answering turn. In this work, we extend beyond the initial work with the following threefold contributions: (i) we use a hierarchical attention network (HAN) architecture in constructing the deception detection model and classify four types of implicatures with non-verbal and pragmatic behavior cues (e.g. backchannel, pause, the change of pitch) in the one model, (ii) we investigate the effectiveness and robustness of multiple speech/language behavior cues consisting of acoustic-prosodic features, semantics, implicatures, and pragmatics for deception detection, and (iii) the proposed framework performs a fusion of turn-level acoustic features and transcripts, and word-level transcripts including non-verbal behavior and pragmatics behavior annotations. The proposed model achieves 80.61% unweighted average recall (UAR) on detecting deception.

The further analyses reveal several insights. Firstly, we observe the same findings in acoustic and conversational temporal dynamics feature set as [20]. Secondly, the truth-tellers have a higher proportion of complications than the liars, which is the same observation as the previous study [22] that was conducted on the English native speakers. Instead, the proportion of common knowledge in liars' behaviors is higher than the truth-tellers. Lastly, the BERT embedding is an effective indicator to detect deceptions. The rest of the paper is organized as follows: database introduction, methodology including introduction of various feature sets, experiments, and conclusions and future work.

II. THE DDDM DATABASE

We evaluate the proposed framework on the Daily Deceptive Dialogues corpus of Mandarin (DDDM) [27] collected at the National Tsing Hua University, Taiwan. It contains 27.2 h of audio recordings of dyadic interaction of Mandarin native speakers with 96 unique speakers (gender is balanced). All participants are paired into dyad over 48 sessions with ages ranging from 20 to 25. There are 7126 sentence-like utterances marked manually.

The DDDM is based on dialog game settings. All subjects are asked to discuss a set of three questions (topics) about their daily life. Three questions are as follows. "Have you ever attended any ball games or competed in ball games?", "Have you ever attended or participated in any concerts?", and "Have you ever attended or performed in any club achievement presentation?". The participants' main goals are to deceive the interlocutors in their answers to one or two of the three questions.

In this work, we follow the same setting as the previous work [20] that pool utterances into QA pair turns. The interrogator tends to ask a series of questions to elicit answers from the deceiver to help recognize the deceptive/truthful statements for each topic. We exclude those interrogators' turns where there is no corresponding answering pair. In other words, we only use complete QA pair turns which means that one questioning turn corresponds to one answering turn (shown in Fig. 2); each turn could have multiple utterances. Then, we convert 7126 sentence-like utterances into 2764 QA pair turns. Finally, DDDM has 283 "question(topic)-level" conversational data samples, and the maximum length of answering turns over all sessions is 40. The evaluation of all experiments uses 10-fold cross-validation (48 sessions are split into 10 folds).

III. RESEARCH METHODOLOGY

Figure 1 illustrates the proposed deception detector model. This is a multi-task learning architecture with two tasks. One is to model non-verbal and pragmatic behaviors at the word-level for implicatures classification, and the other is to detect deceptive-truthful statements. More specifically, pragmatic behaviors contain the deceivers' non-verbal, verbal cues, and abnormal behaviors, and implicature behaviors describe deceivers' metaphors. Besides, there are two phases in the framework: in-phase (i): we model deceivers'

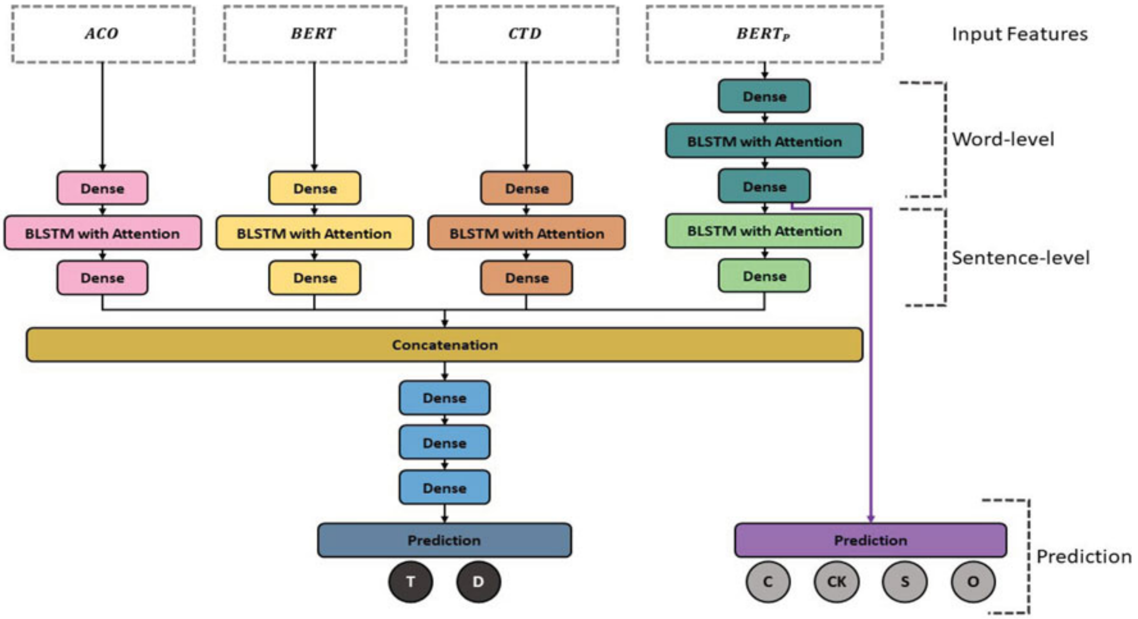

Fig. 1. The overview of the proposed deception detection model. *ACO*, *BERT*, and *CTD* indicate the turn-level acoustic-prosodic features, textual embeddings extracted by BERT pretrained model, and conversational temporal dynamics features proposed in [20], respectively. Further, $BERT_P$ means the word-level textual embeddings extracted by BERT pretrained model, and it also includes non-verbal and pragmatic behavior information. Besides, *C*, *CK*, *S*, and *O* represent complications, common knowledge details, self-handicapping strategies, and others (none of the above them), respectively.

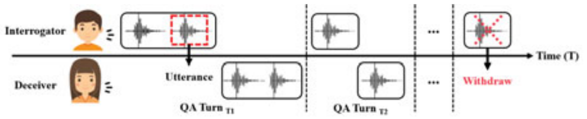

Fig. 2. An illustration of questioning-answering (QA) pair turns. We only use "complete" QA pair turns and exclude that questioning turns if there is no corresponding answering turns. Each turn could have multiple utterances.

behaviors using acoustic features, conversational temporal dynamics, and BERT embedding, and we use non-verbal and pragmatic behaviors to recognize the four-class implicatures; in-phase (ii): we concatenate all the outputs of the last dense layer (before the prediction layer) from every model in the first phase with late fusion,and fine-tune the embeddings with three additional dense layers. Each building block is based on the structure proposed in [28], which consists of an initial dense (fully-connected) layer, then a bidirectional long short-term memory (BLSTM) network with an attention mechanism, and a final dense layer (BLSTM-DNN). Moreover, the HAN structure is inspired by [29], which contains a word-level bidirectional Gated Recurrent Unit (GRU) encoder with an attention mechanism and a sentence-level bidirectional GRU encoder with an attention mechanism. We replace GRU with LSTM in this study for better performance and a fair comparison with the previous work on the same corpus.

A) Deception Detection Framework

In Fig. 1, the model is built based on a BLSTM-DNN within the HAN structure, which is similar to previous studies [28, 29]. It is set up as a multi-task (two tasks) framework. One is to model the deceivers' acoustic and pragmatic behaviors in word-level text for implicatures classification. The other is to model acoustics, textual information, and conversational temporal dynamics for deception detection. Finally, we freeze every trained model and fuse them with late fusion, and then we fine-tune the whole framework with three additional dense layers for deception detection.

Besides, the unit for deceiver's features shown in Fig. 1 is the answering-turn, which includes all of the utterances from the deceiver within a complete QA pair. However, the conversational temporal dynamics feature set is computed based on dyadic turn-taking behaviors. In this work, the deceiver is regarded as the target speaker, and the following sections will introduce all of the feature sets and the proposed BLSTM-DNN within HAN structure in detail.

1) Turn-level Acoustic-Prosodic Features

We extract the same turn-level acoustic-prosodic features as the previous work [20]. It contains 988 acoustic features per utterance computed using the openSMILE toolbox. Specifically, it contains low-level descriptors (LLDs) such as fundamental frequency (pitch) and its envelope, intensity, loudness, 12 MFCC, probability of voicing, eight line spectral frequencies, zero-crossing rate, and delta regression coefficients. Then, the following functionals[1] are further applied on these extracted LLDs with their delta coefficients to generate the final feature vector, denoted by *ACO*. The detailed information can be accessed in the link.

[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): skewness, (19): stddev

2) Turn-level Conversational Temporal Dynamics

The conversational temporal dynamics feature set is firstly proposed in the previous study [20], which is inspired by the prior literature on conversational analyses [21, 22, 30]. It contains 20-dimensional temporal features computed on conversational utterances in each QA pair, denoted by *CTD*. The set includes features such as silence-duration ratio, utterance-duration ratio, silence-utterance ratio, backchannel times, etc. All features are normalized with respect to each speaker using z-score normalization. A brief description is below:

- **Duration:** the total turn duration (d) of interrogator's questioning turn or deceiver's answering turn, denoted as Int_d and Dec_d.
- **Duration difference:** the durational difference between each of the interrogator's and deceiver's turns within a QA pair turn. It is calculated as $Dec_d - Int_d$, and $Int_d - Dec_d$.
- **Duration addition:** the sum of Dec_d and Dec_d.
- **Duration ratio:** the ratio between Res_d and Int_d, and Int_d and Dec_d.
- **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.
- **Hesitation time (h):** the difference between the onset time of the deceiver's utterance and the offset time of the interrogator's utterance, denoted as Dec_h.
- **Backchannel times (bt):** the number of times that a subject interrupts his/her interacting partner, denoted as Int_{bt} and Dec_{bt}.
- **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}.

3) Textual Embeddings

To investigate whether language use would improve automatic deception detection, we first recruit six annotators to transcribe the DDDM database. All of them are native Mandarin Chinese speakers, and each has gone through training and instructed by two of our research members. The annotators also receive an explanation about the DDDM to understand a high-level overview idea of the DDDM. There are a total of 48 conversations that are assigned randomly to six annotators. Furthermore, to ensure the quality of the transcripts, all of the transcripts are reviewed by two of our members. Asides from standard transcription, we ask annotators to further mark non-verbal sounds, such as stammer, laugh, sigh, cough, and unknown and also label pitch change patterns, e.g. the pitch is increasing or decreasing.

Moreover, we use two types of representations from deceivers' answering turn, i.e. turn-level and word-level embeddings. Inspired by [18], we compute a turn-level representation by utilizing BERT, which is a neural language model proposed by [31] to be used as language representation. The BERT model leverages a large amount of plain text data publicly available on the web and is trained in unsupervised objective functions. Specifically, we use BERT-Base Chinese version to extract 768-dimensional encoding, and we exclude all punctuations before feeding the transcribed text to the BERT encoder, denoted by *BERT*. Also, the Chinese BERT-based model works at the character level, so we use it to extract a word-level embedding. Firstly, we perform word segmentation with CKIP-style Chinese NLP tools [32], and then we use BERT to extract word-level embeddings for predicting four types of implicatures. Notice that we put back all punctuation into transcripts for capturing non-verbal and pragmatic behaviors in Table 1 (described in Section A.4), denoted by $BERT_P$. Finally, all features are normalized to each speaker using z-score normalization.

Table 1. The number and annotation of each acoustic and pragmatic behavior.

Meaning	Notation	Number
Smooth pause	,	14 127
Abnormal pause	○	5866
Stuttering	*	1526
Laugh	&	789
Sigh	$	14
Cough	%	24
Long pitch	~	2766
Increase pitch	>	168
Decrease pitch	<	24
Overlap	()	3884
Taiwanese	⌈⌋	10
Indecipherable sounds	◎	194
Unjudgeable sounds	?	84

4) Pragmatics and Implicature Features

There are two phases for the process of annotation collection and labeling. The first phase is to label the basic acoustic features and pragmatics as we expect a higher occurrence of certain features when deception occurs. Subsequently, we decide to focus on the implicatures given in a conversation for the second phase of the labeling. Normally, we tend to expect truth-tellers to be able to give more detailed information than liar since it takes imagination and higher cognitive effort for liars to make up something that never truly happen. Instead of measuring the total amount of information in a conversation, Vrij et al. [22] provide a new criterion that is easier to measure and more effective in detecting deception. Specifically, it involves calculating the proportion of three different categories of complications. The remaining parts of the section will explain the details in the DDDM database.

Non-verbal Acoustic and Pragmatic Features

The non-verbal acoustic and pragmatic features are labeled during the transcribing process by six annotators. They are instructed to mark the features listed in Table 1.

For example, when the subjects pause during a conversation, the annotator then has to distinguish whether it is a smooth pause or an abnormal pause. If it is a smooth pause, the annotator mark a "," in the transcript. Then, we regard the labels in Table 1 as segment delimiter set when doing word segmentation using the CKIP-style Chinese NLP tools. Finally, we use BERT-Base Chinese version pre-trained model to extract 768-dimension representation (notice that we put the labels back to the original position in the sentence before BERT encoding). We expect that the model can learn non-verbal cues in human spoken dialogs.

Types of Implicatures

For the labeling implicatures, the three-class implicatures that we label in theDDDM are mainly based on [22] with some modifications due to language and culture differences between Chinese and English. During this labeling process, we first translate the parts that are classified as details in the transcripts. The definition of implicatures in the non-essentials responses from speakers. Next, implicatures are subdivided into three categories: *Complication (C)*, *Common knowledge details (CK)*, and *Self-handicapping strategies (S)*. Also, the left sentences are regarded as another category, *Others (C)*.

For the annotation process of implicatures, there is a single annotator that finishes all the transcripts to increase the consistency of the data annotation. If the annotator is uncertain about the labels, the two researchers would discuss with the annotator but not directly change the annotator's labeling. The annotator marks the features by examining the transcription instead of listening to the whole audio recordings. Finally, there are 450 C, 56 CK, 22 S, and 2088 O in total. Three types of implicatures are described in the following.

- **Complication (C):** A complication refers to details associated with personal experience or knowledge learned from any personal experience. The DDDM includes subjects of college students from the university (NTHU) and the nearby school (NCTU). The three topics/questions that are assigned to each subject during the deception game are about general activities and experiences of the average college student. As a result, the contents of the collected conversations have a high degree of similarity. We are then able to strictly define whether certain contents are personally related or not. For instance, scores of department border cups, professional knowledge about instruments, and detailed process of any events held by different clubs, etc., are regarded as personal experiences.
- **Common knowledge details (CK):** A common knowledge detail refers to details associated with common experiences or general knowledge about events. Especially, general knowledge is defined as knowledge that every NTHU and NCTU student should know. For example, the final exam week for the semester, the location of buildings on the campus, the school bus stop locations, Meichu games between NTHU and NCTU, and so on. As for the common experiences, it is defined as experiences that the NTHU and NCTU students should know, e.g. studying at the library, eating McDonald's at the cafeteria, taking the school bus at the school, to name a few.
- **Self-handicapping strategies (S):** A self-handicapping strategy refers to explicit or implicit justification as to why the speaker is not able to provide information [22]. Notice that if someone simply states that he/she forgets about something, then it does not classify as self-handicapping strategies. The speaker has to give a direct or indirect excuse for not being able to provide more information about the situation.

IV. EXPERIMENTS

A) EXPERIMENTAL SETUP

The HAN-like structures consist of three levels of modeling: word-level, sentence-level, and the proposed deception-level modeling. We follow [29, 33] to use word sequence encoder, word-level attention, sentence encoder, and sentence-level attention from HAN [29], and we reform it for deception detection. The followings will describe each fold in modeling.

1) Three Folds in Modeling

- **Word-level Modeling:** This modeling captures the relationship between acoustic-pragmatic behaviors and implicatures, and we directly model acoustic-pragmatics using textual embeddings to predict the four types of implicatures through BLSTM-DNN. The model learns from word-level inputs to recognize the four-class implicatures, and the recognition results are shown in Table 2. The maximum word length of all answering turns, which is done by word segmentation with CKIP-style Chinese NLP tools [32], is 598. We further utilize zero-padding to fix the length when the length of answering turns is less than 598.
- **Sentence-level Modeling:** There are four types of models based on different inputs in this stage. Firstly, we freeze the weights of the model for implicatures classification and regard it as a word-level representation encoder. This encoder will encode each answering-turn into sentence-level representation, and we train an additional BLSTM-DNN for deception recognition. On the other hand, the left three BLSTM-DNN models are trained for deception detection from other feature sets, such as acoustic-prosodic features and conversational temporal dynamics. In the second stage, the maximum length of the question-answering turn pair is 40, and the length of turns which are less than 40 will also be zero-padded. Table 2 shows the results of using different feature sets, and the model trained with BERT features achieves the best results without fusing with other models.
- **The Proposed Deception Modeling:** We freeze weights of all models on the deception detection task and concatenate their final dense layer's outputs as the input to an additional three-layer feed-forward neural network to perform a late fusion of different feature sets. Besides, we compare the performance of these models with the

Table 2. Results (%) on the DDDM database presented with metrics of unweighted accuracy recall (UAR), weighted-F1 score, and macro-precision.

Fusion method	Feature set	Overall (UAR)	Deception (UAR)	Truth (UAR)	Weighted-F1	Macro-precision
-	Human	55.55	40.52	70.59	54.71	56.11
	ACO [20]	70.31	68.94	71.67	70.03	70.53
	BERT	74.06	74.27	73.85	73.17	76.00
	$BERT_P$	65.29	67.86	62.72	64.38	66.86
	CTD [20]	66.02	77.91	54.14	64.87	68.37
Early fusion	*ACO* + *CTD* [20]	74.71	74.89	74.53	74.39	75.52
	ACO + *BERT*	76.22	81.36	71.09	75.42	78.16
	BERT + *CTD*	74.69	81.97	67.41	74.47	75.65
	ACO + *BERT* + *CTD*	77.00	79.78	74.21	76.63	77.44
Late fusion	*ACO* + *BERT*	78.28	77.65	78.91	77.46	79.08
	ACO + $BERT_P$	72.11	72.40	71.82	71.73	72.52
	BERT + $BERT_P$	75.59	78.94	72.23	75.41	76.22
	ACO + *CTD*	74.38	72.00	76.77	73.49	75.47
	BERT + *CTD*	76.76	77.91	75.60	75.86	78.92
	$BERT_P$ + *CTD*	71.68	87.16	56.19	69.49	75.55
	ACO + $BERT_P$ + *CTD*	74.34	77.10	71.58	72.86	76.58
	ACO + *BERT* + *CTD*	80.05	83.25	76.85	79.20	81.18
	BERT + $BERT_P$ + *CTD*	76.51	86.11	66.90	75.74	78.45
	ACO + *BERT* + $BERT_P$	78.88	79.56	78.19	78.15	79.81
	ACO + *BERT* + $BERT_P$ + *CTD*	**80.61**	**80.34**	**80.87**	**79.95**	**81.37**

feature-level fusion method (early fusion), and all the results are shown in Table 2. Notice that the result of the model with $BERT_P$ does not show in the early fusion because the word-level characteristic is different from other turn-level features. Finally, the proposed model achieves an 80.61% UAR.

2) Evaluation and Experimental Parameters

In the word-level training stage, the number of hidden nodes in the BLSTM and the dense layer are 64 and 128, respectively. On the other hand, in the other training stages, the number of hidden nodes in the BLSTM and the dense layer are 8 and 16. The evaluation of all experiments uses 10-fold cross-validation with weighted-F1, macro-precision, and the metric of UAR, which is equal to macro-recall. Moreover, we set batch size 64, learning rate 0.0005 with ADAMAX optimizer [34], and cross-entropy as our loss function. The number of epochs is chosen with early stopping criteria in all experiments on the validation set, and our proposed framework is implemented using the PyTorch toolkit [35].

B) Experimental Results and Analyses

1) Analyses of Model Performance

Table 2 shows all model performance, and the model with all feature sets (late fusion) achieves the best overall UAR of 80.61%. Also, we found that *CTD* feature set has higher performance for deception-class detection compared with other features when examining performances obtained using a single feature set. On the other hand, the *BERT* feature set is better at recognizing truth-class detection, and it also achieves great accuracy on deception-class. Besides, Table 3 shows a summary of the results for four-class implicatures recognition using the model trained with $BERT_P$. Although the data distribution is very unbalanced, these annotations contain rich information about personal experience according to [22]. Finally, we obtain 62.03% UAR on the implicatures classification task. We use this network as the encoder to extract representations, i.e. characterizing vocal behaviors of telling details, and we obtain deception detection task performance of 65.29% UAR.

To compare with the state-of-the-art (SOTA) performance on DDDM [20], we train the models with an early fusion of different feature sets, which is the same setting as the previous work [20]. The proposed model surpasses the best result in [20] by 5.9% absolute in the DDDM. Because the word-level feature, $BERT_P$, can not be included in the early fusion, we do not show its result in Table 2. In addition, we observe that the BERTembedding is more effective in improving the performance of deception detection than other feature sets. We also find that the model trained with *ACO*, *BERT*, and *CTD* features in late fusion has a competitive performance comparing with the proposed model (with all features). However, the proposed model has slightly performed better on truthful class, and this result fits our expectation because the implicatures information can help untrained people catch the truth-tellers from the liars. This finding is similar to previous psychologists' studies [22, 23, 36].

2) Analyses of Input Features

We follow Vrij and Vrij [23] to calculate the proportion of each implicature class. They investigated the implicature behaviors affected by cultural differences (Russian, Koreans, and Hispanic) between truth-tellers and liars. The proportion of *common knowledge (CK)* is equal to the number of *CK* divided by the total number of three-class implicatures (the number of *self-handicapping strategies (S)* plus the number of *common knowledge (CK)* details plus the number of *complications (C)*).

Here, we conduct a similar study of implicature on the DDDM. Table 4 shows a summary of the number and the

Table 3. Results and the data distribution of the four-class implicatures recognition on the DDDM database. We present metrics of unweighted accuracy recall (UAR), weighted-F1, and macro-precision (%). *C*, *CK*, *S*, and *O* represent complications, common knowledge details, self-handicapping strategies, and others (none of the above them), respectively.

Feature	Overall (UAR)	C (UAR)	CK (UAR)	S (UAR)	O (UAR)	Weighted-F1	Macro-precision
$BERT_P$	62.03	73.12	43.37	47.00	84.62	85.57	46.85
Number	-	450	56	22	2088	-	-

Table 4. The proportion of implicature classes by calculating the number of each implicature divided by the total number of three types of implicatures (the number of self-handicapping strategies plus the number of common knowledge details plus the number of complications).

	Truth-tellers		Deception-tellers	
Implicature type	Number	Proportion	Number	Proportion
C	222	0.851	215	0.846
CK	28	0.107	28	0.11
S	11	0.042	11	0.043

Table 5. The Welch's *T*-test results between truthful and deceptive answering turns in the three feature sets. A feature's value and the number of features all are smaller than 0.05 (if a feature's *p*-value is <0.01, it is marked by *.)

Feature set	Number	Number*	Feature
ACO	57	21	Please see Table 6
CTD	5	0	Int_{ud}, Int_{su}, Int_{d}/Dec_{d}, Int_{us}, Int_{st}
BERT	179	91	-

Table 6. Welch's *T*-test between truthful and deceptive responses in acoustic features. A feature's value and the number of features all are <0.05 (if a feature's *p*-value is <0.01, it is marked by *).

Feature	Functional (in footnote of page 3)
$\Delta MFCC_{8th}$	2, 3, 4, 8*, 9*, 10, 12, 14, 16, 17, 19*
$MFCC_{8th}$	1, 2, 3, 8*, 9*, 12*, 14*, 17, 18, 19*
$MFCC_{6th}$	3, 8*, 9*, 10, 12, 17*, 19*
$\Delta MFCC_{6th}$	8*, 9*, 10, 17, 19*
Loudness	3*, 4*, 8*, 9, 19*
$MFCC_{9th}$	2, 8, 9, 18, 19
$\Delta Foenv$	2*, 5, 18
$\Delta MFCC_{7th}$	4, 16, 19
$\Delta MFCC_{12th}$	10, 17
ΔZCR	6, 7
$\Delta LspFreq_{7th}$	14
$\Delta VoiceProb$	10
$MFCC_{12th}$	5
$MFCC_{2th}$	4

proportion of each implicature. Interestingly, we have similar findings as to the work by Vrij [23]. According to their studies [22, 23, 36], truth-tellers tend to give more complications than deceivers because deceivers prefer to make their stories simple, and deceivers are more prone to express common knowledge details in their explanations than truth-tellers sincedeceivers have no personal experiences to describe. Moreover, deceivers tend to have more self-handicapping strategies in their stories than truth-tellers because deceivers do not want to provide too many details.

Besides, we perform a statistical Welch's *T*-test between truthful and deceptive answering responses with regard to *ACO*, *BERT*, and *CTD* following a similar study framework as [20]. Table 5 shows the features whose *p*-values are smaller than 0.05 and 0.01. In terms of *ACO* and *CTD*, we obtain similar findings as [20],there are 57 dimensions of acoustic parameters where *p*-values obtained are smaller than (<) 0.05, and 21 features among them are smaller than (<) 0.01. Specifically, $MFCC_{8th}$, $MFCC_{6th}$, and their first derivatives are useful indicators for detecting deceptions, which is the similar finding in the previous works on an English corpus [11] and the Mandarin Chinese corpus [20].

On the other hand, *CTD* obtained from the interrogator's behaviors (like Int_{ud}) are important indicators in showing whether the deceiver is telling the truth or not. Also, we listen to the recordings of DDDM and observe that the interrogator would often ask more complicated questions and spend more time thinking about what the next question they want to ask those segments when the deceivers are producing lies. This particular finding is quite interesting as we find that the "Human" labeled accuracy is relatively low on identifying deceptive events. However, interrogators' behaviors (maybe unconscious) would directly indicate whether he/she is indeed being given a truthful/deceptive answer. Besides, there are 179 dimensions of *BERT* where *p*-values obtained are < 0.05, and 91 features among them are < 0.01. That is, *BERT* are an important feature representation to help provide discriminatory power in differentiating whether the deceiver is telling the truth or not.

V. CONCLUSIONS AND FUTURE WORK

In this study, we have investigated the integration of a suite of speech and language features in characterizing the spoken dialog content to help improve the performance of the automatic deception detection framework. In comparison to previous studies on automatic deception detection, we provide a much wider and exhaustive range of insights and findings on human deception behaviors as manifested in the spoken language content, such as the proportional of implicatures, conversational temporal dynamics, and non-verbal/pragmatics behaviors. These features are then fed into the proposed multitask and multistage BLSTM-based with HAN to perform deception detection. The proposed model is evaluated on a recently collected large Mandarin Chinese database, DDDM, and achieves SOTA performance of 80.61% UAR. Besides, thefour-class implicatures classifier achieves 62.03% accuracy with only textual information. Throughout this work, we have shown that (i) the proportional of complications of truth-tellers and the proportional of common knowledge of liars have the same trend as the previous research [22] done on English native speakers; (ii) textual information is an important indicator to detect deception behaviors comparing to other feature sets. In the immediate work, inspired by our promising results obtained text information, one of the key efforts is to integrate an automatic speech recognition to investigate the robustness of theframework. We would further like to integrate syntactic information and personality traits scores [37] to enhance the recognition power of the deception detection model. Furthermore, we also want to extend our work to 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 in the DDDM.

ACKNOWLEDGEMENT

The work is supported in part by funding from the Ministry of Science and Technology, Taiwan. We thank the contribution of Yi-Tong Wu, Chih-Hsiang Huang, and Ching-Yun Hsu for collecting the audio recordings and various annotations in the DDDM database. We further thank the anonymous reviewers and Ching-Yun Hsu for their valuable comments.

REFERENCES

[1] Sarkadi, S.: Deception, in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, ser. IJCAI'18. AAAI Press, 2018, 5781–5782.

[2] Clementson, D.E.: Truth bias and partisan bias in political deception detection. *J. Lang. Soc. Psychol.*, **37** (4) (2018), 407–430. [Online]. Available: https://doi.org/10.1177/0261927X17744004.

[3] Conroy, N.K.; Rubin, V.L.; Chen, Y.: Automatic deception detection: methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, **52** (1) (2015), 1–4. [Online]. Available: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2015.145052010082.

[4] Vaccari, C.; Chadwick, A.: Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, **6** (1) (2020), 2056305120903408. [Online]. Available: https://doi.org/10.1177/2056305120903408.

[5] Grazioli, S.; Jarvenpaa, S.L.: Consumer and business deception on the Internet: content analysis of documentary evidence. *Int. J. Electron. Commerce*, **7** (4) (2003), 93–118.

[6] Triandis, H.C. *et al.*: Culture and deception in business negotiations: a multilevel analysis. *Int. J. Cross Cultural Manag.*, **1** (1) (2001), 73–90. [Online]. Available: https://doi.org/10.1177/147059580111008.

[7] Vrij, D.A.; Graham, S.: Individual differences between liars and the ability to detect lies. *Expert Evid.*, **5** (4) (1997), 144–148.

[8] Stiff, J.B.; Kim, H.J.; Ramesh, C.N.: Truth biases and aroused suspicion in relational deception. *Communic. Res.*, **19** (3) (1992), 326–345. [Online]. Available: https://doi.org/10.1177/009365092019003002.

[9] Swol, L.M.V.; Braun, M.T.; Kolb, M.R.: Deception, detection, demeanor, and truth bias in face-to-face and computer-mediated communication. *Communic. Res.*, **42** (8) (2015), 1116–1142. [Online]. Available: https://doi.org/10.1177/0093650213485785.

[10] Levitan, S.I.; Maredia, A.; Hirschberg, J.: 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, June 2018, pp. 1941–1950. [Online]. Available: https://www.aclweb.org/anthology/N18-1176.

[11] Levitan, S.I.; Maredia, A.; Hirschberg, J.: Acoustic-prosodic indicators of deception and trust in interview dialogues, in *Proceeding of the Interspeech 2018*, 2018, 416–LPAGE420. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-2443.

[12] Venkatesh, S.; Ramachandra, R.; Bours, P.: Robust algorithm for multimodal deception detection, in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. Los Alamitos, CA, USA: IEEE Computer Society, March 2019, 534–537. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/MIPR.2019.00108.

[13] Pérez-Rosas, V.; Abouelenien, M.; Mihalcea, R.; Burzo, M.: 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, 59–LPAGE66. [Online]. Available: https://doi.org/10.1145/2818346.2820758.

[14] Aune, R.; Waters, L.L.: Cultural differences in deception: motivations to deceive in Samoans and North Americans. *Int. J. Intercult. Relat.*, **18** (2) (1994), 159–172. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0147176794900264.

[15] Rubin, V.L.: TALIP perspectives, guest editorial commentary: pragmatic and cultural considerations for deception detection in Asian languages. *ACM Trans. Asian Lang. Inf. Process.*, **13** (2) (2014), Article No: 10. https://doi.org/10.1145/2605292.

[16] Thannoon, H.H.; Ali, W.H.; Hashim, I.A.: Design and implementation of deception detection system based on reliable facial expression. *J. Eng. Appl. Sci.*, **14** (15) (2019), 5002–5011. [Online]. Available: https://medwelljournals.com/abstract/?doi=jeasci.2019.5002.5011.

[17] Liu, X.; Hancock, J.; Zhang, G.; Xu, R.; Markowitz, D.; Bazarova, N.: Exploring linguistic features for deception detection in unstructured text, in *Hawaii International Conference on System Sciences*, YEAR2012. [Online]. Available: https://sml.stanford.edu/pubs/2012/exploring-linguistic-features-for-deception-detection-in-unstructured-text/.

[18] Gröndahl, T.; Asokan, N.: Text analysis in adversarial settings: does deception leave a stylistic trace?. *ACM Comput. Surv.*, **52** (3) (2019), Article No: 45. https://doi.org/10.1145/3310331.

[19] Wu, Z.; Singh, B.; Davis, L.; Subrahmanian, V.: Deception detection in videos, 2018. [Online]. Available: https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16926.

[20] Chou, H.; Liu, Y.; Lee, C.: 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)*, November 2019, 1044–1050.

[21] Vrij, A.; Hartwig, M.; Granhag, P.A.: Reading lies: nonverbal communication and deception. *Annu. Rev. Psychol.*, **70** (1) (2019), 295–317. pMID: 30609913. [Online]. Available: https://doi.org/10.1146/annurev-psych-010418-103135.

[22] Vrij, A.; Leal, S.; Jupe, L.; Harvey, A.: Within-subjects verbal lie detection measures: a comparison between total detail and proportion of complications. *Legal Criminol. Psychol.*, **23** (2) (2018), 265–279. [Online]. Available: https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/lcrp.12126.

[23] Vrij, A.; Vrij, S.: Complications travel: a cross-cultural comparison of the proportion of complications as a verbal cue to deceit. *J. Invest. Psychol. Offender Profiling*, **17** (1) (2020), 3–16. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jip.1538.

[24] Vrij, A. *et al.*: 'Please tell me all you remember': a comparison between British and Arab interviewees' free narrative performance and its implications for lie detectionSEP. *Psychiatry Psychol. Law*, (2020), 1–14.

[25] Kontogianni, F.; Hope, L.; Taylor, P.J.; Vrij, A.; Gabbert, F.: Tell me more about this. . .': an examination of the efficacy of follow-up open questions following an initial account. *Appl. Cogn. Psychol.*, **34** (5) (2020), 972–983. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.3675.

[26] Vrij, A.; Leal, S.: Proportion of complications in interpreter-absent and interpreter-present interviews. *Psychiatry Psychol. Law*, **27** (2020), 155–164.

[27] Huang, C.-H.; Chou, H.-C.; Wu, Y.-T.; Lee, C.-C.; Liu, Y.-W.: Acoustic indicators of deception in Mandarin daily conversations recorded from an interactive game, in *Proceeding of the Interspeech 2019*, 2019, 1731–1735. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2216.

[28] Mirsamadi, S.; Barsoum, E.; Zhang, C.: Automatic speech emotion recognition using recurrent neural networks with local attention, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, 2227–2231.

[29] Singhania, S.; Fernandez, N.; Rao, S.: 3HAN: a deep neural network for fake news detection, in *Neural Information Processing*, Cham: Springer International Publishing, 2017, 572–581.

[30] Benus, S.; Gravano, A.; Hirschberg, J.: Pragmatic aspects of temporal accommodation in turn-taking. *J. Pragmat.*, **43** (12) (2011), 3001–3027. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378216611001469.

[31] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K.: 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, June 2019, 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423.

[32] Li, P.-H.; Fu, T.-J.; Ma, W.-Y.: Why attention? Analyze BiLSTM deficiency and its remedies in the case of NER, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **34** (05) (2020), 8236–8244. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6338.

[33] Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.J.; Hovy, E.H.: Hierarchical attention networks for document classification, in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, 1480–1489. [Online]. Available: http://aclweb.org/anthology/N/N16/N16-1174.pdf.

[34] Kingma, D.P.; Ba, J.: 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*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980.

[35] Paszke, A. *et al.*: PyTorch: an imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019, 8026–8037. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

[36] Vrij, A. *et al.*: Using the model statement to elicit information and cues to deceit in interpreter-based interviews. *Acta. Psychol. (Amst)*, **177** (2017), 44–53. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001691816301998.

[37] Chou, H.; Lee, C.: 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)*, December 2020.

Huang-Cheng Chou received the B.S. degree in electrical engineering (EE) from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2016. He is currently working toward a Ph.D. degree with the EE Department, NTHU, Hsinchu, Taiwan. His research interests are in automatic deception detection and automatic emotion recognition. He won the Best Regular Paper Award on the APSIPA ASC 2019 and Merry Electroacoustics Thesis Award 2020 (top 12). He was the recipient of the Graduate Students Study Abroad Program sponsored by the Taiwan Ministry of Science and Technology (MOST), the NTHU President's Scholarship, the FUJI Xerox Research Award, and the travel grant sponsored by ACII 2017, the Foundation for the Advancement of Outstanding Scholarship, and the Association for Computational Linguistics and Chinese Language Processing. He is a Co-Author on the recipient of the 2019 MOST Futuretek Breakthrough Award. He is a Student Member of the APSIPA, ISCA, and IEEE Signal Processing Society.

Yi-Wen Liu received his B.S. degree in Electrical Engineering (EE) from National Taiwan University in 1996, and M.S. and Ph.D. degrees in EE from Stanford University in 2000 and 2006, respectively. He was a post-doctoral researcher at Boys Town National Research Hospital, Omaha, USA from 2006 to 2010. In 2010, he joined the Department of Electrical Engineering at National Tsing Hua University (NTHU) and is currently a full professor. His research focuses on hearing, speech, and audio signal processing. He received NTHU Outstanding Teaching Award in 2014, 2017, and 2020 and was hence granted a distinguished professorship by the university. He is a member of IEEE Signal Processing Society and the Acoustical Society of America.

Chi-Chun Lee is an Associate Professor at the Department of Electrical Engineering of the National Tsing Hua University (NTHU), Taiwan. He received his B.S. degree and Ph.D. degree in Electrical Engineering from the University of Southern California (USC), USA in 2007 and 2012. His research interests are in speech and language, affective multimedia, and health analytics. He is an IEEE senior member. He is an associate editor for IEEE Transaction on Affective Computing (2020), IEEE Transaction on Multimedia (2019-2020), and a TPC member for APSIPA IVM and MLDA. He is a coauthor on the best paper award/finalist in INTERSPEECH 2008, 2010, 2018, IEEE EMBC 2018, 2019, 2020, and APSIPA ASC 2019. He is the recipient of the Foundation of Outstanding Scholar's Young Innovator Award, the CIEE Outstanding Young Electrical Engineer Awar, the IICM K. T. Li Young Researcher Award (2020), the MOST Futuretek Breakthrough Award (2018, 2019).