
Hierarchical Tensor Fusion Network for Deception Handling Negotiation Dialog Model

Nguyen The Tung
NAIST, Japan
nguyen.tung.np5@is.naist.jp

Koichiro Yoshino
NAIST, Japan
PRESTO, JST, Japan
koichiro@is.naist.jp

Sakriani Sakti
NAIST, Japan
AIP, RIKEN, Japan
s-sakti@is.naist.jp

Satoshi Nakamura
NAIST, Japan
AIP, RIKEN, Japan
s-nakamura@is.naist.jp

Abstract

Handling deceptive information of conversation partner is effective to lead a successful negotiation. However, capturing deceptive information requires to consider multi-modal information of conversation partner. In this paper, we propose a multi-modal tensor fusion network for deception detection to be used in deception handling dialogue manager of negotiation dialog. Our method can be viewed as a combination of two existing integration methods of multi-modal features, hierarchical fusion, and tensor fusion. Experimental results showed that the proposed hierarchical tensor fusion network outperformed existing fusion methods in the accuracy of deception detection. Moreover, the negotiation dialog manager that utilizes the proposed method for deception detection achieved good performance in terms of the system's action selection accuracy.

1 Introduction

Using deception is a common tactic in negotiation dialog to reach an outcome that is the most beneficial to the interlocutor. In the research area of negotiation dialog systems, it is known that handling deceptive information in dialog strategies has a substantial impact on the success of the negotiation [1]. Thus, it is critical to build an accurate deception detection module when constructing negotiation dialog systems.

According to existing studies, acoustic factors, such as pitch, intensity, and speaking rate, can vary when someone is being deceptive [2]; thus, acoustic features are generally used for automatic deception detection. On the other hand, some other studies suggested that visual clues (facial expressions) are important in identifying deceptions; it is reported that there are relations between facial expressions and the deception labels [3, 4]. Based on the results from these studies, it is expected that incorporating facial features to deception detection based on acoustic clues has the potential to increase detection accuracy.

When we exploit features from several modalities, there are many possible ways to integrate these features. Neural networks currently achieve good scores on a variety of tasks; thus, we focus on feature integration methods in neural network based approaches. One prospective approach is hierarchical integration [5], which is proposed for emotion recognition. This architecture has a hidden layer that connects with both acoustic and facial features but does not concatenate these feature vectors as a simple feature integration method, under the assumption that acoustic and facial features

have different levels of abstractions. Another candidate is utilizing a tensor fusion network (TFN) that composes a tensor of the acoustic features and facial features. TFN is reported as the best method for improving the accuracy of face recognition and sentiment analysis tasks, because it can consider the outer product of different type features explicitly [6, 7]. Thus, we expected TFN is able to model the important features contributing to the improvement of deception detection accuracy.

Despite being successful in various tasks, those combination methods still have some drawbacks. The hierarchical fusion combines modalities using concatenation, which is highly inefficient and complex for training. On the other hand, tensor fusion network still treats all modalities to be at the same level of abstraction. Based on these existing studies, we propose a new type of neural network for combining multi-modal features, called the hierarchical tensor fusion network (Hierarchical TFN). Our proposal is a combination of a hierarchical and a tensor fusion network and has the advantages of both of these methods: balancing the feature abstraction level by a hierarchical structure and explicitly combining different types of features by outer product.

Experimental results indicated that the hierarchical tensor fusion achieved the best score, outperformed the existing methods. We used the predicted labels from different classification models as inputs of an reinforcement-learning-based dialog manager of negotiation dialog [8]. The results also indicated that the performance of the action selection of the dialog manager when using the predicted deception labels from the hierarchical tensor fusion based classifiers achieved a good score.

2 Related work

There are a number of existing studies that investigated useful information for automatic deception detection: acoustic-prosodic features [2], gestures from hand and head motions [9], and lexical features [10]. When we plan to use these features in deception detection for negotiation dialog systems, it is difficult to expect the contribution of lexical features because the system needs to label each utterance even if the utterance contains a few words (e.g., answers to yes/no questions). Most deception detection by using linguistic features is conducted on the paragraph level; therefore, we do not include the linguistic features in our deception detection module.

A multi-modal approach is widely used for a variety of tasks such as emotion recognition or sentiment analysis. In contrast to the single-modal, a multi-modal approach utilizes features from different (usually two or three) modalities for its respective task. Currently, this approach is being used more for the task of deception detection because it is reported that there is some improvement by including visual (gestures, head movements, etc.) or acoustic features (pitch and power) along with lexical clues, which is traditionally used for deception detection [11, 12, 13]. Multi-task learning of personality recognition and deception detection was also proposed by using multi-modal features [14].

A major drawback of current studies about multi-modal deception detection is that the majority of them are based on a simple concatenation of features or a linear interpolation of several classification results. In contrast, for other similar tasks such as emotion recognition, neural networks are widely used, and it is reported that hierarchical fusion [5] or a tensor fusion network [6, 7] outperforms traditional integration methods, especially for emotion recognition. The study by Amiriparian [15] showed that emotions play an important role in deceptive expressions; thus, we expected that hierarchical fusion or TFN would also perform well for the task of deception detection.

3 Deception detection based on acoustic and facial features in several integration methods

This section describes different methods of modalities combination. The classification model we used in this study is multi-layer perceptron (MLP), which is a fully-connected feed-forward neural network with one hidden layer generally used for classification problems. The output contains two neurons with softmax activation to determine the probability of deception. All the fusion methods used in our study are an extension of this network.

3.1 Early and late fusion

Early and late fusions of features are common ways to integrate multiple modalities. Network architectures of early and late fusion methods are shown in Figure 1. With the early fusion, we simply concatenate the vectors that contain acoustic and visual features into one single vector and then feed

it to the MLP network. This is the most widely used combination method in previous studies for deception detection.

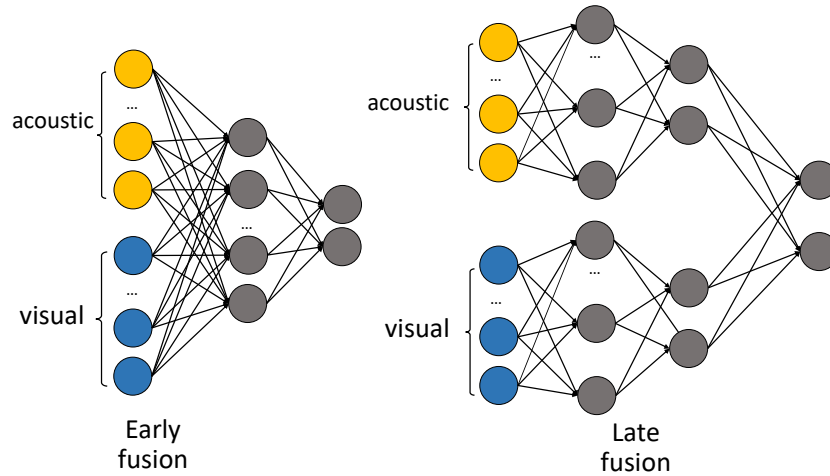


Figure 1: *Early and late fusion architecture.*

For the late fusion model, the network can be viewed as a combination of two MLPs that is similar to early fusion; one network takes visual features as input, and the other network uses acoustic features. The outputs of these two subnetworks are then fully connected to a final output layer with two neurons. Models are trained separately among the two subnetworks and the network that connects them together into the output; thus, results from two networks are considered in equal weight.

3.2 Hierarchical fusion

Hierarchical fusion is a method to combine different modalities proposed by [5] in a work for emotion recognition. Figure 2 describes the structure of this fusion method for the case of fusing visual and acoustic modalities.

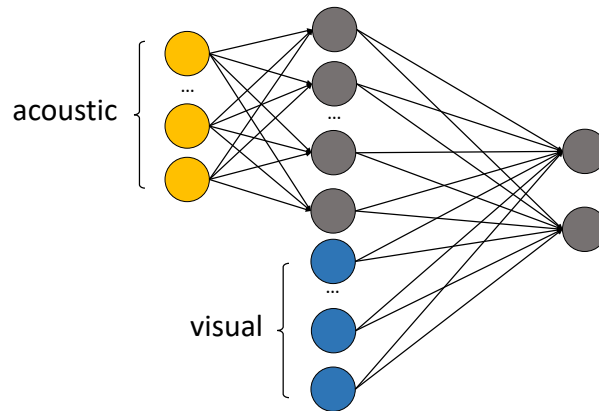


Figure 2: *Hierarchical fusion architecture.*

As the name suggests, the architecture of this fusion model resembles a hierarchy graph with the layers of network equivalent to levels in a hierarchy. In this figure, the vector of acoustic features is fed into the input layer, which is fully connected to a hidden layer. The vector containing visual features is concatenated with this hidden layer. The resultant vector is fully connected to an output layer. [5] argued that different modalities may describe data at different timescales or have different

levels of abstraction, thus features from different modalities should be put into different layers of the network. In particular, the features that describe data at a larger timescale and are more abstract are used in higher levels as shown in Figure 2. Detailed parameter settings are described in Section 4. We used similar numbers of parameters in different models to investigate the performances of model architectures.

3.3 Tensor Fusion Network

The tensor fusion method is another approach for combining different modalities in the task of sentiment analysis [7] and face recognition [6]. [7] discussed that when combining multiple modalities, a neural network needs to learn about both intra-modality and inter-modality feature interactions. From Figure 1, we can see that the late fusion network cannot learn inter-modality interactions. On the other hand, the early-fusion network learns both kinds of interactions simultaneously, so training is difficult. With TFN, learning of intra-modality and inter-modality interactions is separated, making the training process easier. Another benefit of TFN is that the representation of inter-modality interactions is given explicitly to the network in the form of the outer product (tensor), thus reducing the complexity of training. Because of these reasons, TFN is expected to work better than early and late fusion methods. Both studies [7, 6] observed that the tensor fusion outperformed early and late fusions in their respective tasks.

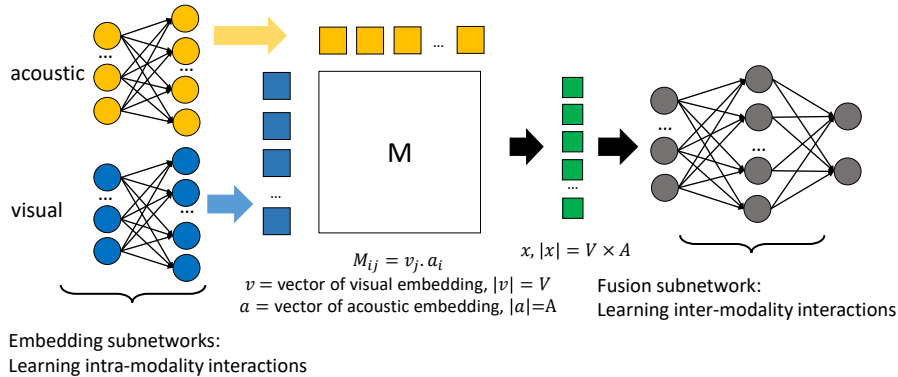


Figure 3: *Tensor fusion architecture.*

A TFN contains two kinds of subnetworks in its structure. The first one is the embedding subnetwork, which performs learning of intra-modality interactions. The outputs of those subnetworks are embedding vectors for each modality. Next, we perform outer production of the embedding vectors (visual and acoustic). The reason we use outer product to combine the vectors is that it can represent all the interactions between each feature from visual and acoustic modalities. The result is a matrix $M : V \times A$ (V and A are the size of the visual and acoustic embedding vectors, respectively), which is then flattened into a vector x with size $V \times A$. We feed this x into the fusion subnetwork (which is a MLP), which has the role of learning about the inter-modality interactions. Previous studies reported that the decomposition of the tensor contributed to classification accuracies; however, we did not observe any improvement by the decomposition. Therefore, we show the results of not using decomposition in our experiments.

3.4 Hierarchical Tensor Fusion Network

The structure of our proposed network is shown in Figure 4. As can be seen from this figure, the hierarchical TFN structure is similar to a hierarchical network, but multi-modality fusion is performed by using the outer product (same as TFN) instead of concatenating. Our proposed method’s advantages over a hierarchical network is similar to that of TFN over early fusion thanks to the use of the outer product for fusion.

The hierarchical TFN also resembles a TFN structure but raw acoustic features are used as inputs of tensor fusion instead of an acoustic embedding vector as we can see in TFN. A strong benefit over TFN that the proposed method has is reduction of the complexity of network structure and the

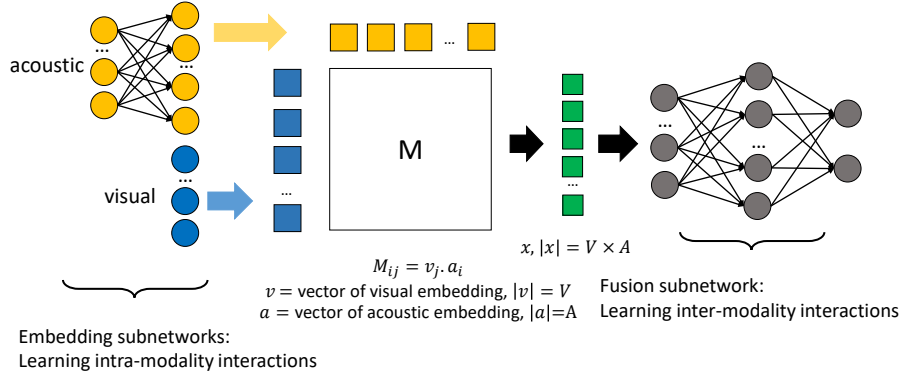


Figure 4: Hierarchical tensor fusion architecture.

number of parameters. In some tasks (such as emotion recognition or deception detection), there can be a modality whose intra-modality interactions are not as beneficial as those of the other modalities. In such situation, using a deep structured network to learn about these intra-modality interactions is superfluous. Therefore, by removing the embedding subnetwork for such modality, the whole system can focus on more important feature interactions and train a better model. In our case, we empirically found out that visual feature interactions is less important than those of acoustic features. Thus, a vector containing raw visual features is used directly for fusion, as shown in Figure 4.

4 Experiments

4.1 Data

The dataset we used for deception detection includes two types of data. The first one is taken from the Real-life Trial dataset [12]. We split the video into segments (as each segment contains a single utterance) to suit our research purpose. If the video was annotated with a lie label in the original dataset, then we assign lie labels to all segments from this video. We manually checked the segmented videos and found out that many of them have low quality (blurry video or the speaker not facing the camera). Therefore, we decided to filter out bad segments by using confidence score provided from the OpenFace toolkit [16]. In particular, if a video does not contain any frame with confidence score of face tracking higher than 0.85 then it is automatically removed. After this process, from the trial dataset, we have a total of 245 utterances; 105 of them are deceptive, and 140 of them are truthful. The second data consists of recorded videos of health consultation dialogs, each dialog was carried out by two participants [8]. The deception labels of each utterances in this dataset were manually annotated by the participants. The second dataset contains 844 truthful and 177 deceptive utterances. In total, our deception detection dataset used in the experiment includes 1265 utterances.

We performed our experiment of deception detection using 4-fold cross-validation. When checking the real-life trial dataset [11], we found out that many of the videos are taken from the same trial recording and were assigned the same label (lie or truth). It means there is a chance that the classifiers learn to predict recordings instead of deception labels, if we randomly split the data to train, development, and test sets. To avoid such problem, we separate the samples in our dataset by the recording that they belong to, from 70 recordings of the original dataset to five portions. Detailed data separation is shown in Table 1.

In previous studies [13, 11, 2], samples were chosen so that the ratio of lie/truth is balance (1:1). We followed the same configuration and allocated the samples in such a way that development and test set have ratio of lie/truth close to 1:1. Particularly, for each split in our cross-validation setup, we pick out two partitions (1-2, 2-3, 3-4, and 4-1) to be used as development and test set; the remaining three partitions are used as training set. In partition 5, the number of honest samples exceeds the number of lie samples by a large margin, therefore, we did not use this partition for development and testing. We used over-sampling to achieve 1:1 ratio between lie and truth samples used for training.

Table 1: Dataset partitions.

	# recordings	Recording ID	# lie	# honest
Partition 1	3	1,2,13	66	77
Partition 2	11	3,4,14,16,17,18, 19,24,26,27,35	59	76
Partition 3	13	10,23,25,28,29, 30,31,32,33,34, 36,37,38	62	79
Partition 4	4	9,15,17,23	60	77
Partition 5	39	the remaining	38	671

4.2 Features

We used the OpenFace toolkit [16] to extract facial features. We were able to extract 14 face action unit (AU) regressions and 6 AU classification values as well as head position, and head direction parameters for each frame by using this toolkit. These values were then normalized and discretized into five different levels of intensity to be used as features for deception detection. Acoustic features were extracted from audio files using the OpenSMILE toolkit [17]. We used the Interspeech 2009 (IS09) emotion challenge standard feature-set as our acoustic features. These features were also used in previous studies of the deception detection task [13, 18]. In conclusion, for each segment sample in our dataset, we were able to extract 78 visual and 384 acoustic features.

4.3 Results

4.3.1 Results of deception detection

In the first experiment, single-visual, single-acoustic, and early fusion models all have two hidden layers, with the first layer have 256 units. The hierarchical fusion model contains two hidden layers. The first layer is fully-connected with the input layer from acoustic modality. This hidden layer has 256 units and is concatenated with the input visual features vector. The resultant concatenation fully connects with the second hidden layer, which is in turn fully-connected with the output. For TFN models, the embedding subnetworks have one hidden layer which has the same number of units as the output embedding vector (32 units). Similar to that from the original work [7], we use ReLU as the activation function for the embedding subnetwork. The fusion subnetwork of TFN has one hidden layer. We did not augment the embedding vectors with 1 since our empirical experiment showed no improvement. In other words, the TFN model used in this experiment is equivalent to $TFN_{bimodal}$ of visual and acoustic modalities from the prior work [7]. With hierarchical TFN, the embedding subnetwork has one hidden layer (256 units), and the output has 32 units. Similar to TFN model, the fusion subnetwork of hierarchical TFN also contains one hidden layer. All the models were trained

Table 2: Results of deception detection.

Model	# layers	# parameters	Accuracy	Precision	Recall	F1-score
single acoustic	4	162,900	53.78%	0.4747	0.5000	0.4870
single visual	4	164,352	49.28%	0.4095	0.3525	0.3879
multi early	4	169,872	53.42%	0.4603	0.3566	0.4018
multi late	4	164,972	54.68%	0.4794	0.3811	0.4247
multi hierarchical	4	168,192	53.78%	0.4733	0.4713	0.4723
multi TFN	5	163,528	50.36%	0.4216	0.3525	0.3839
multi hierarchical TFN	5	166,448	58.63%	0.5304	0.5000	0.5148

using the Adam optimizer [19] with a softmax cross entropy loss function. We used development loss to tune up the learning rate; the remaining hyper-parameters are the default setting of Chainer¹. We trained models using mini-batch, with a batch size of 16. The learning rate decreased by 10% at every epoch. The loss on the development set was used to determine the point to stop; if we did not see improvement of the development loss for 100 epochs, the training was stopped. All the models are configured so that they have similar number of parameters, ranging from 163,000 to 170,000.

¹<https://chainer.org/>

Table 2 summarizes the parameters and performance of different models in the deception detection task. In this table, “single” refers to the models that use only one modality (visual or acoustic), and “multi” refers to the models that use both modalities. Early, late, hierarchical, and TFN indicate their integration methods. The numbers are averaged from 4-fold cross-validation results. Accuracy is measured for both labels (truthful and deceptive). Precision, recall, and F1-score are measured for the deceptive label.

From these results, it is clear that the proposed hierarchical TFN has the highest overall performance, outperforming both hierarchical fusion ($p < 0.05$) and TFN ($p < 0.05$). In particular, we can see large improvement in terms of precision and F1-score compared to the other models.

We can see that the single visual model performs a bit worse than single acoustic model ($p \approx 0.105$) in term of accuracy, while precision, recall, and F1-score are much lower. This difference can be explained by how we extracted features from raw data. Acoustic features are extracted from raw audio at 100 frames per second (fps) while visual features are extracted at only 30 frames per second (since all videos are recorded at 30 fps). Therefore, with the same spoken utterance, acoustic features can capture more refined information. Thus, we can expect that contribution to performance from the visual modality is less than the acoustic one. Another reason for the disparity in performance between visual and acoustic modality is the nature of recordings in our deception dataset. Within every recording, the speaker knows that their statement will be assessed for honesty and thus, the speaker tries to conceal their facial expressions as much as possible to avoid getting caught lying. Hence, detecting deception from visual clues (facial expressions) is not trivial.

4.3.2 Relationship between deception detection accuracies and numbers of parameters

In this experiment, we assess the effect of network depth on deception detection performance of hierarchical structures. We measured the accuracy when changing the number of hidden layer in hierarchical fusion and hierarchical TFN models. With hierarchical fusion, the number of hidden layer refers to the number of layers between the resultant concatenation and the output layer. For hierarchical TFN, number of hidden layers refers to the hidden layers of the fusion subnetwork. The 1-layer models are taken directly from the previous experiment. With the 1-layer models as a base structure, we add new hidden layers (each has 256 units) to construct the 2-layer and 3-layer models.

Figure 5 illustrates the relations between the accuracies and the layer numbers for both the hierarchical architecture and the hierarchical tensor fusion network architecture. We can see that there is no significant gain by increasing number of hidden layers from one, two, or three. It indicates that a deep structure not always contributes the accuracy of deception detection, especially for a small dataset as we used. However, it is difficult to prepare a large-scale dataset for deception detection.

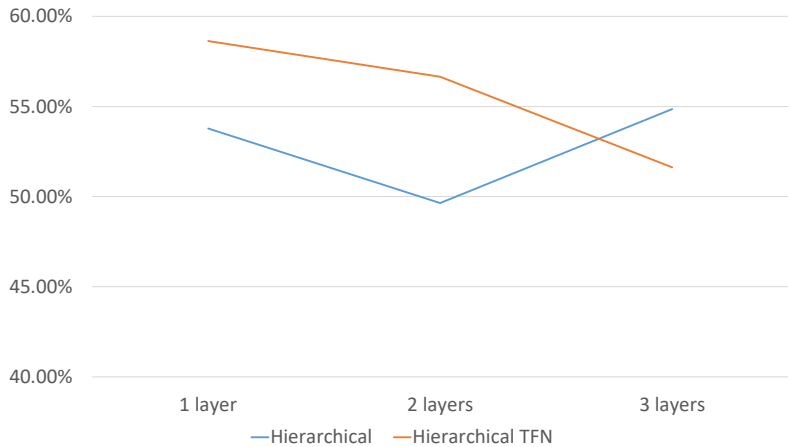


Figure 5: Effect of network depth on detection performance.

Table 3: Accuracy of dialog acts selection when using different deception labels result.

Deception labels used for dialog management	DA accuracy
chance rate deception	65.69%
gold-label deception	80.31%
single visual prediction	70.15%
single acoustic prediction	66.22%
multi early prediction	66.48%
multi late prediction	68.58%
multi hierarchical prediction	69.10%
multi TFN prediction	69.66%
multi hierarchical TFN prediction	71.20%

4.4 Dialog management based on predicted deception labels

In this experiment, we used predicted labels from deception detection models for a negotiation system that decides output dialog acts (DA) on the basis of the user’s deception information (whether the user is lying or not). The dialog tactics of the system should change in accordance with the user deception; thus, we used a Q-learning based dialog manager that chooses the system’s dialog act by using the deceptive information of users. The dialog manager we used is similar to the one described in [8]. Particularly, we modeled the dialog decision process using Partially Observable Markov Decision Process (POMDP) and constructed dialog state from user DA and user deception information. Our Q-learning based dialog manager was trained by interaction with a user simulator. This experiment was also conducted using the health consultation data from [8], where the system acts as the health consultant and persuades a human user to adopt a more healthy lifestyle. The system can choose one from three available dialog acts as response to the user. Each utterance (from use or system) was assign one DA label.

We measured the system’s performance by DA selection accuracies, which refers to the precision of the system’s chosen dialog acts against reference actions that were chosen by a human when given the same user input utterance. All of the results showed in Table 3 used the policy trained with gold-labels of deception and the dialog act. Similar to the previous experiment, we found that the negotiation system achieves the highest DA selection accuracy when using deception labels from the proposed hierarchical TFN model. This result indicates that the proposed method contributes not only to deception detection but also helps the dialog system achieve high performance as well.

5 Conclusions

In this study, we proposed a new method called hierarchical tensor fusion network for combining visual and acoustic modalities for the task of deception detection in negotiation dialog. Experimental results indicated that the hierarchical tensor fusion model has the best performance, outperforming the existing approaches used in previous studies (hierarchical and tensor fusion). We also investigated in experiments about the DA selection accuracy of a negotiation dialog system when using output labels from multiple deception detection methods and found out that the dialog system achieves high performance when using labels from the hierarchical TFN model.

In the future, our first focus is to collect or augment more data to fully exploit this powerful fusion model and build a full negotiation dialog system that can handle deceptive information. In addition, we would also like to apply our method to other similar tasks such as emotion recognition or sentiment analysis to confirm the findings in this study.

6 Acknowledgments

Part of this work was supported by JSPS KAKENHI Grant Number JP17H06101.

References

- [1] Ioannis Efstathiou and Oliver Lemon, “Learning to manage risks in non-cooperative dialogues,” in *Proceedings of the 18th Workshop on the Semantics and Pragmatics of Dialogue (SemDIAL)*, Edinburgh, Scotland, U.K., September 2014, pp. 173–175.
- [2] Julia Hirschberg, Stefan Benus, Jason M Brenier, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, Laura Michaelis, et al., “Distinguishing deceptive from non-deceptive speech,” in *Interspeech*, 2005, pp. 1833–1836.
- [3] Paul Ekman and Wallace V Friesen, “Detecting deception from the body or face.,” *Journal of personality and Social Psychology*, vol. 29, no. 3, pp. 288, 1974.
- [4] Paul Ekman, “Darwin, deception, and facial expression,” *Annals of the New York Academy of Sciences*, vol. 1000, no. 1, pp. 205–221, 2003.
- [5] Leimin Tian, Johanna Moore, and Catherine Lai, “Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features,” in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 565–572.
- [6] Guosheng Hu, Yang Hua, Yang Yuan, Zhihong Zhang, Zheng Lu, Sankha S Mukherjee, Timothy M Hospedales, Neil Martin Robertson, and Yongxin Yang, “Attribute-enhanced face recognition with neural tensor fusion networks.,” in *Proceedings of ICCV 2017*, 2017.
- [7] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, “Tensor fusion network for multimodal sentiment analysis,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 1103–1114, Association for Computational Linguistics.
- [8] The Tung Nguyen, Koichiro Yoshino, Sakriani Sakti, and Satoshi Nakamura, “Impact of deception information on negotiation dialog management.,” *IWSDS*, 2018.
- [9] Thomas O Meservy, Matthew L Jensen, John Kruse, Judee K Burgoon, Jay F Nunamaker, Douglas P Twitchell, Gabriel Tsechpenakis, and Dimitris N Metaxas, “Deception detection through automatic, unobtrusive analysis of nonverbal behavior,” *IEEE Intelligent Systems*, vol. 20, no. 5, pp. 36–43, 2005.
- [10] Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards, “Lying words: Predicting deception from linguistic styles,” *Personality and social psychology bulletin*, vol. 29, no. 5, pp. 665–675, 2003.
- [11] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo, “Deception detection using real-life trial data,” in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 59–66.
- [12] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, Yao Xiao, CJ Linton, and Mihai Burzo, “Verbal and nonverbal clues for real-life deception detection.,” in *EMNLP*, 2015, pp. 2336–2346.
- [13] Hosomi Naoki, Sakti Sakriani, Koichiro Yoshino, and Satoshi Nakamura, “Deception detection and analysis in spoken dialogues based on fasttext,” *Proc. APSIPA 2018*, 2018.
- [14] Guozhen An, Sarah Ita Levitan, Julia Hirschberg, and Rivka Levitan, “Deep personality recognition for deception detection,” *Proc. Interspeech 2018*, pp. 421–425, 2018.
- [15] Shahin Amiriparian, Jouni Pohjalainen, Erik Marchi, Sergey Pugachevskiy, and Björn W Schuller, “Is deception emotional? an emotion-driven predictive approach.,” in *INTERSPEECH*, 2016, pp. 2011–2015.
- [16] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency, “Openface: an open source facial behavior analysis toolkit,” in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–10.

- [17] Florian Eyben, Martin Wöllmer, and Björn Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.
- [18] Gideon Mendels, Sarah Ita Levitan, Kai-Zhan Lee, and Julia Hirschberg, “Hybrid acoustic-lexical deep learning approach for deception detection,” *Proc. Interspeech 2017*, pp. 1472–1476, 2017.
- [19] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *ICLR 2015*, 2015.