# Deception Detection and Analysis in Spoken Dialogues based on FastText

Naoki Hosomi*, Sakriani Sakti*†, Koichiro Yoshino*†‡, Satoshi Nakamura*†
* Nara Institute of Science and Technology, Japan
† RIKEN, Center for Advanced Intelligence Project AIP
‡ PRESTO, Japan Science and Technology Agency
E-mail: {hosomi.naoki.hg6, ssakti, koichiro, s-nakamura}@is.naist.jp Tel/Fax: +81-743-72-5264

***Abstract*—Detecting deception is complicated for humans even though it often happens in human communications. In contrast, machines can capture small features to achieve accurate deception-detection, which is difficult for humans. Classifiers based on supervised learning make it possible to analyze effective features for deception-detection by giving positive and negative samples of deception to the classifier. FastText is one accurate classifier for a variety of classification problems, sentiment analysis, or the tagging of sentences, all of which use the distributed representation of features. We constructed a deception detector for dialogue utterances by giving labels of deception to FastText. We also combined acoustic features for deception-detection and analyzed the deception-detection results. The resultant detector achieved significantly higher accuracy than deception-detection by humans.**

## I. Introduction

Deception is an act that intentionally causes another a person or persons to hold a false belief. This social behavior is done by most of us on a daily basis, based on the inevitable conflicts of interest in human interaction. When deception happens, it violates the (usually tacit) agreement between the two parties of information exchanges and thus represents a misuse of and a threat to communication. As an example of such a situation, imagine a speaker who profits from and a listener who suffers a disadvantage in job interviews [1], [2], [3]. If fraud can be accurately detected, we can avoid such unreasonable disadvantages.

However, previous researchers argued that the detection of deception by humans is difficult. Bond et al. carried out a meta-analysis of research on more than 200 different previous fraud detections and reported that the average correct deception-detection rate by a person without special training was about 54% [4]. Levine et al. experimentally detected deception by changing the proportion of truth that was included in messages presented to participants. Their results revealed that accuracy depends on the specific portion of the presented truth [5]. In other words, accuracy does not substantially exceed a chance level even when people are striving to detect deceptions. One various claim as to why we cannot detect fraud is that humans suffer from biases. For example, we tend to judge speech as valid regardless whether it is actually true (i.e., truth-bias) [6]. Perhaps humans cannot recognize specific phenomena during deception, and another cause of our difficulty detecting fraud is that we often focus on clues that are unrelated to deception.

In this study, we constructed a classifier that detects deception at a higher performance level than humans. Classifiers based on supervised learning make it possible to analyze effective features for deception-detection by giving positive and negative samples of deception to them. FastText is one accurate classifier for a variety of classification problems, sentiment analysis, or tagging of sentences that use the distributed representation of features. We constructed a deception detector for dialogue utterances by giving labels of deception to FastText. We also combined acoustic features for deception-detection and analyzed the deception-detection results. We investigated the features that are effective for detection and compared the performances of our system with humans.

## II. Related Work

Regarding fraud detection methods, there are reports that claim that high detection performance can be obtained by measuring human physiological responses using polygraphs, fMRIs, etc. [7]. However, it is difficult to detect fraud during free communication by connecting complicated instruments to the identification target or extracting responses through a particular question procedure. Therefore, in this research, we discuss fraud detection using language and speech features that do not require such special systems.

A previous study that correlated linguistic features with deceptive behavior [8] classified deception and truthfulness from real-world sources, criminal narratives, interrogations, and legal testimony. Another study's approach utilized the non-verbal behavior data of users from social media to detect multiple account identity deceptions [9]. Torres et al. [10] performed a study of glottal waveform features for deceptive speech classification, and Zhou et al. [11] constructed deception detection from speech signal using relevance vector machine and non-linear dynamics features. But these studies only used either verbal or non-verbal cues. Perez-Rosas et al. [12] utilized both verbal and non-verbal features to build a multimodal system to detect deception in real-life settings.

Some existing research on deception-detection has been done by a statistical learning method with spoken language from human conversations. Hirschberg et al. created the Columbia/SRI/Colorado (CSC) Corpus of deceptive speech for training and testing. It contains about seven hours of speaker speech including deception with Standard American

English [7]. Then based on this corpus, they constructed deception detectors using the Ripper rule-induction classifier that utilizes various features. A classifier based on acoustic, prosodic, and speaker dependent features achieved the best accuracy, which was improved by about 6% compared to a chance level. Recently, Levitan et al. created a corpus for English speakers who have both Standard English and Mandarin as their mother tongue. They used RandomForest as a discriminator and its acoustics and prosodic features as well as the characteristics of the speakers and achieved accuracy about 10% higher than a chance level [13]. But the study did not perform deception-detection with a deep learning framework.

Recently, Mendels et al. reported a fully connected neural network that learned in a speech and bidirectional LSTM that learned distributed expressions created by GloVe [14] By combining models, the F1-scores reached 0.64: precision = 0.67, recall = 0.61 [15]. However, most studies did not analyze or compare their results with human performances. In contrast, this study discusses the performance differences between machines and humans.

## III. Proposed Method

### A. Features

A work by Amiriparian et al. argued that emotion scores estimated from speech are effective for deception-detection from a corpus that contains human-agent conversation [16]. Therefore, in this work, we used 384-dimensional acoustic features from the INTERSPEECH 2009 (IS09) emotion recognition challenge [17] using the openSMILE toolkit [18].

We also use word embedding as a lexical feature and propose a method using fastText [19] architecture, whose structure resembles the Continuous Bag-of-Words (CBOW) model of Word2Vec [20], where the middle word is replaced by a label. Thus it constructs word embedding using the information of sentence labels and simultaneously learns the weight of the classifier.

Figure 1 illustrates the fastText architecture with lexical $N$ n-gram features $x_1, ..., x_N$. These lexical features are then embedded into $e_1, ..., e_N$ and averaged to form hidden variable $m$, which is in turn fed to a linear classifier. Hidden variable $m$ is a text representation that can be reused for other tasks. Here the softmax function computes the probability distribution over the predefined classes, where $y$ is the output label.

Joulin et al. argued that this model achieved high performance in such tasks of text classification as sentiment analysis [19]. In our case, since we used a text classification task for deception-detection, we expected to optimize the word embedding learned by fastText and the deception corpus for deception-detection.

### B. Classifier

As a classifier, we utilized multilayer perceptron (MLP), which is a kind of feedforward neural network that is comprised of multiple layers from Rosenblatt's perceptron [21]. In this work as an activation function, we used sigmoid function:

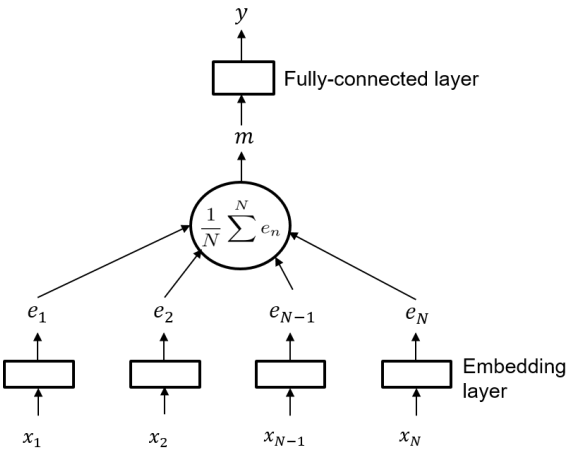


Fig. 1. Architecture of fastText

$$h(x) = \frac{1}{1 + \exp(-x)}. \quad (1)$$

The classifier's hyperparameters are tuned by Bayesian optimization, and the expected improvement is based on maximizing the accuracy for the validation of the dataset used in our experiment. In this experiment, the range of the hyperparameters was the number of layers (2 to 5), the number of units (128, 192, 256), and the dropout rate (0.1 to 0.5) of the hidden layer of the multilayer perceptron.

We define $n$ as the number of samples for $K$ samples, comparing output $y_n$ with given label $t_n$. While the classifier learns, we can calculate the loss function by binary cross entropy:

$$E = -\sum_{n=1}^{k} \{t_n \log y_n + (1 - t_n) \log(1 - y_n)\}. \quad (2)$$

The same classifier is used for both the acoustic and linguistic features. The overall architecture of the proposed model is shown in Fig. 2.

The classifier learns by minimizing loss. In our proposed method, the loss (Eq. 2) is monitored during the learning, which is stopped after three epochs according to the loss decreasing.

## IV. Experimental Set-Up

### A. Dataset

TABLE I
DIALOG EXAMPLES FROM CSC CORPUS

| Utterance | Label |
|---|---|
| Well, yeah, there's a chance. | Truth |
| Uh, actually, I did well. Excellent. | Deception |

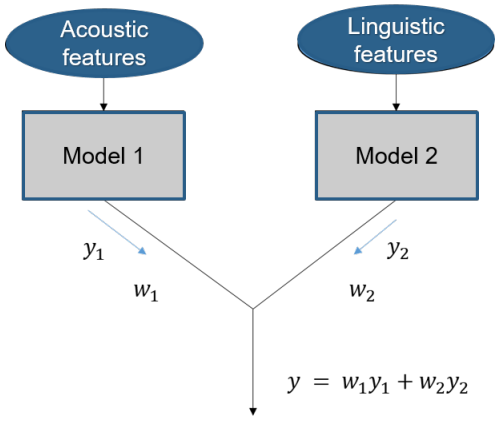


Fig. 2. Overall architecture of proposed model

We used CSC Deceptive speech[1] to make our experiment's dataset that originally consisted of about 32 hours of audio interviews with 32 native speakers (16 males and 16 females) of Standard American English. We utilized every utterance in the interviews as a dataset for training and testing; since it is hard to use every utterance because the dataset contains too many short utterances, extracting useful features is difficult. Since predicting utterance labels from extremely short utterances is also difficult, we used utterances longer than five words.

We called our experiment, which used randomly selected utterances, a random test. We chose this strategy to exclusively focus on the relationships between the deception labels and the features that can be extracted from a single utterance: in other words, ignoring the effect of context-level features. In this experiment, the test data were 100 utterances (truth 50 and deception 50), and the training data were 4000 utterances (truth 2000 and deception 2000) randomly selected from the entire speech in the corpus of the interviewees. As we mentioned in the introduction, since there is truth-bias, we adjusted the dataset's ratio at a 50% chance level.

### B. Human Participants

In this research, we also experimentally checked whether humans and classifiers can detect deception. Our participants read and listened to deceptive and non-deceptive (truth) speeches and then predicted whether those speeches were deceptive or true.

The participants were non-native English speakers whose English proficiency TOEIC scores were 730 or higher (Mean = 861.7, SD = 117.7): six male graduate students, three from Japan and three from other countries.

After our participants read and listened to an utterance from our test data, they predicted its label. To identify the tendency whether utterances are deceptive or true, participants freely confirmed the utterances and the labels of the training data before they took the test of deception detection.

[1]https://catalog.ldc.upenn.edu/LDC2013S09

### C. Training of Proposed Classifier

We used 90% of the training data (1800 true and 1800 false utterances) as the training set and 10% (200 true utterances and 200 lies) of the data for tuning the parameters. Then we predicted the labels of the test set.

In this experiment, we didn't use the dialogue's context information, and the judgments were only made from utterances. Also, we used a Theano (version 0.9.0) backend Keras (version 1.2.1) for all the implementation of the classifiers. We used batch size 10 and epoch number 10 for learning both models and Adam as an optimization algorithm.

Based on the input's language feature, a bag of bigrams was input into the discriminator, and a 30-dimensional distributed representation was constructed. Based on the acoustic features, in the discriminator 10% of the units were randomly removed at each layer of the middle layer and learning was performed. For a combination of acoustic and lexical classifiers, we set the weights for lexical classifier $w1 = 0.6$ and acoustic classifier $w2 = 0.4$ based on the highest accuracy with the validation set.

### D. Metrics

Our evaluation used the following formula with deceptive utterances as a positive example (P) and true utterances as a negative example (N) as deception-detection metrics.

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{FN + TP} \tag{5}$$

$$F - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{6}$$

## V. Experiment Results and Discussion

Regarding our experimental results, Table II shows the human deception-detection results and Table III shows the detection results by the proposed classifier.

For experiments on human deception-detection, we calculated Fleiss's Kappa coefficient to investigate the degree of agreement of the predicted labels for utterances among experimental participants. In general, when the Kappa coefficient is 0.6 or more, the degree of coincidence of the prediction is considered high, but in our experiment it was 0.16, causing variations in the labels predicted by the experiment collaborators. For Table II for the task of a chance level of 50%, the average human accuracy performance was about 51%. A one-sided binomial test on this accuracy showed no significant difference compared with a chance level ($p > 0.05$).

For deception-detection with our proposed classifier, the best accuracy was 10% higher than a chance level when we combined a classifier using acoustic and lexical features. After carrying out a one-sided binomial test on the detection results obtained only by listening to the speech with the highest

TABLE II
HUMAN PERFORMANCE

| | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Speech | 0.515 | 0.524 | 0.370 | 0.414 |
| Text | 0.510 | 0.515 | 0.387 | 0.425 |
| Speech+Text | 0.512 | 0.498 | 0.360 | 0.405 |

TABLE III
CLASSIFIER PERFORMANCE

| | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Acoustic Feat. | 0.580 | 0.577 | 0.600 | 0.588 |
| Lexical Feat. | 0.62 | 0.630 | 0.580 | 0.604 |
| Acoustic + Lexical Feat. | 0.640 | 0.667 | 0.560 | 0.609 |

accuracy in human detection, we confirmed that deception-detection with significantly higher precision than humans was possible by the proposed method ($p < 0.05$). We also confirmed that the F-measure obtained by the proposed method significantly outperformed humans.

We distributed surveys to our human participants after the experiments. The majority answered that both the acoustic and linguistic features were useful for detection; however, we found no significant difference among any experiment conditions (Table II). Some participants said that they made their decisions based on such fillers as "um" or "uh" as well as types of emphasis words, stuttering, intonation, power, and pitch. However, our results indicate that these criteria were not useful. Note that since our experiment was conducted with non-native speakers, we need to verify this result with native speakers.

## VI. CONCLUSIONS

We conducted a comparative verification of deception-detection ability on interview-style speech with humans and statistical learning methods. Our results suggest that the human ability to detect deception is approximately chance level regardless whether a dialogue context was included. Our result also suggests that humans tend to predict an utterance to be true regardless of an actual label. We also confirmed that classifiers by the statistical learning method are more accurate than humans and a chance level. In addition, the classifier used word embeddings and acoustic features that can outperform a scheme that uses only one feature. In the future, we will construct a more accurate classifier based on the history of a conversation. We will also conduct a deception-detection with native English speakers.

## ACKNOWLEDGMENTS

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17K00237 and JP17H00747 as well as JST PRESTO (JPMJPR165B).

## REFERENCES

[1] WP Robinson, A Shepherd, and J Heywood, "Truth, equivocation concealment, and lies in job applications and doctor-patient communication," *Journal of Language and Social Psychology*, vol. 17, no. 2, pp. 149–164, 1998.

[2] Julia Levashina and Michael A Campion, "Measuring faking in the employment interview: development and validation of an interview faking behavior scale.," *Journal of applied psychology*, vol. 92, no. 6, pp. 1638, 2007.

[3] Brent Weiss and Robert S Feldman, "Looking good and lying to do it: Deception as an impression management strategy in job interviews," *Journal of Applied Social Psychology*, vol. 36, no. 4, pp. 1070–1086, 2006.

[4] Charles F Bond Jr and Bella M DePaulo, "Accuracy of deception judgments," *Personality and social psychology Review*, vol. 10, no. 3, pp. 214–234, 2006.

[5] Timothy R Levine, Rachel K Kim, Hee Sun Park, and Mikayla Hughes, "Deception detection accuracy is a predictable linear function of message veracity base-rate: A formal test of park and levine's probability model," *Communication Monographs*, vol. 73, no. 3, pp. 243–260, 2006.

[6] Steven A McCornack and Malcolm R Parks, "Deception detection and relationship development: The other side of trust," *Annals of the International Communication Association*, vol. 9, no. 1, pp. 377–389, 1986.

[7] 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.

[8] Joan Bachenko, Eileen Fitzpatrick, and Michael Schonwetter, "Verification and implementation of language-based deception indicators in civil and criminal narratives," in *Proceedings of the 22nd International Conference on Computational Linguistics*, 2008, pp. 41–48.

[9] Michail Tsikerdekis and Sherali Zeadally, "Multiple account identity deception detection in social media using nonverbal behavior," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 8, pp. 1311–1321, 2014.

[10] Juan Torres, Elliot Moore, and Ernest Bryant, "A study of glottal waveform features for deceptive speech classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 4489–4492.

[11] Yan Zhou, Hemming Zhao, Xinyu Pan, and Li Shang, "Deception detecting from speech signal using relevance vector machine and non-linear dynamics features," *Neurocomputing*, vol. 151, no. 3, pp. 1042–1052, 2015.

[12] Veronica Perez-Rosas, Mohamed Abouelenien, Rada Mihalcea, Yao Xiao, CJ Linton, and Mihai Burzo, "Verbal and nonverbal clues for real-life deception detection," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2336–2346.

[13] Sarah I Levitan, Guzhen An, Mandi Wang, Gideon Mendels, Julia Hirschberg, Michelle Levine, and Andrew Rosenberg, "Cross-cultural production and detection of deception from speech," in *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. ACM, 2015, pp. 1–8.

[14] Jeffrey Pennington, Richard Socher, and Christopher Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[15] 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.

[16] 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.

[17] Björn Schuller, Stefan Steidl, and Anton Batliner, "The interspeech 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

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

[19] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[20] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[21] Frank Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain.," *Psychological review*, vol. 65, no. 6, pp. 386, 1958.