Automatic Detection of Deceptive and Truthful Paralinguistic Information in Speech using Two-Level Machine Learning Model

Velichko A.N.
SPC RAS
Saint-Petersburg, Russia
alena.n.velichko@gmail.com

Karpov A.A.
SPC RAS
Saint-Petersburg, Russia
karpov@iias.spb.su

Abstract

In this work, we present a novel approach to one of computational paralinguistic tasks – automatic detection of deceptive and truthful information in human's speech. This task belongs to the aspects of destructive behaviour and was first presented at the International INTERSPEECH Computational Paralinguistics Challenge ComParE in 2016. The need of contactless method for deception detection follows from the fact that existing contact-based approaches such as polygraphs and lie detectors have multiple restrictions, which significantly limit their usage. Both for training and testing of the proposed models we used two English-language corpora (Deceptive Speech Database and Real-Life Trial Deception Detection Dataset). We extracted tree sets of acoustic features from those audio samples using openSMILE toolkit. The proposed approach includes preprocessing of the extracted acoustic features with the usage of methods for data augmentation and dimensionality reduction of feature space. We have got 1680 speech utterances and 986-dimensional informative feature vector for each utterance. The main part of the proposed approach is two-level recognition model, where the first level includes three models of gradient boosting (Catboost, XGBoost and LightGBM). The second level consists of logistic regression-based model for final prediction on truthfulness or deceptiveness that takes into account predictions from the first level. Using this approach, we have achieved the result of classification in terms of F-score = 85.6%. The proposed approach can be used both independently and as a component of multimodal systems for detection of deceptive and truthful utterances in speech, as well as in systems for detection of a destructive behaviour.

Keywords: speech technology, computational paralinguistics, detection of deceptive and truthful information in speech, machine learning, gradient boosting, multimodal systems

DOI: 10.28995/2075-7182-2021-20-698-704

Автоматическое определение ложной и истинной паралингвистической информации в речи человека с применением двухуровневой модели машинного обучения

Величко А.Н.
СПб ФИЦ РАН
Санкт-Петербург, Россия
alena.n.velichko@gmail.com

Карпов А.А.
СПб ФИЦ РАН
Санкт-Петербург, Россия
karpov@iias.spb.su

Аннотация

В работе предложен подход к решению одной из задач компьютерной паралингвистики – автоматическому определению ложной и истинной информации в речи человека. Данная задача является одним из аспектов деструктивного поведения и впервые была представлена на международных соревнованиях по компьютерной паралингвистике INTERSPEECH ComParE в 2016 году. Необходимость бесконтактного метода определения ложной информации в речи вытекает из того, что существующие контактные подходы, например, полиграфы или детекторы лжи, имеют ряд требований, которые значительно ограничивают их использование. Для обучения и тестирования предложенных моделей нами использовались два корпуса англоязычной речи (Deceptive Speech Database и Real-Life Trial Deception Detection Dataset), из аудиозаписей которых были вычислены три набора акустических признаков посредством программного инструментария openSMILE. Предложенный подход включает предобработку акустических признаков при помощи метода аугментации данных и метода уменьшения размерности признакового пространства, что в итоге позволило получить 1680 речевых высказываний, из которых были вычислен 986-размерный вектор

информативных признаков. Основой предложенного подхода является двухуровневая модель, в которой на первом уровне используются три модели градиентного бустинга (Catboost, XGBoost и LightGBM), а на втором уровне – модель на основе логистической регрессии, которая позволяет выдавать итоговое предсказание о ложности или истинности высказывания на основе предсказаний моделей первого уровня. С использованием такого подхода удалось добиться значения результата определения ложной и истинной информации по показателю F-меры, равного 85,6%. Предложенный подход может использоваться как самостоятельно, так и в качестве компонента многомодальной системы определения ложной и истинной информации в речи или системы определения деструктивного поведения.

Ключевые слова: речевые технологии, компьютерная паралингвистика, определение ложности и истинности информации в речи, машинное обучение, градиентный бустинг, многомодальные системы

1 Introduction

The task of automatic detection of deceptive and truthful information in speech belongs to the field of computational paralinguistics as well as detection of mental, emotional and physical states, detection of different diseases (including COVID-19) by voice, speech and noises etc. Moreover, lie is one of the aspects of destructive behaviour that also includes depression, aggression, etc. Nowadays, with the advancement of the Internet and social networks such destructive behaviour is more common to appear in text format [16]. On the other hand, both in virtual and real life people still use natural speech for communication and it is also a study object. In recent years many groups of researches have presented papers addressing contactless deception detection in speech. The reason is that current contact-based methods have multiple restrictions that refer both to a place and a research subject. The concept of automatic deception detection in speech is based on the hypothesis that telling lies have an impact on increasing stress level and it affects acoustic parameters. Additionally, linguistic, para- and extralinguistic factors (such as different psychophysiological states, pathologies of mentality or mental disorders, and some other diseases) have an impact on phonetic characteristics of speech and even possibility of speech production [15].

There are unimodal and multimodal approaches to detection of deceptive and truthful information in speech. Unimodal systems can be used both by itself and as a part of more complex systems for psychophysiological human states. Multimodal systems for automatic deception detection by speech can significantly improve the quality of classification because of their ability to analyze additional paralinguistic aspects. Those aspects include mimics and gestures (eyebrows movements, lips tension, gaze direction, hands movement etc.), they are informative markers for detection of deceptive or truthful speech utterances. Moreover, analysis of lexical component of speech utterance can allow different markers in speech, for example, uncertainty expressed by particular words, hesitations and interjections. Contactless systems can be used in such areas as banking (for example, loan granting), law enforcement (for example, polygraph tests, preventing of “telephone terrorism” etc.).

The first time the task of contactless detection of deceptive and truthful information was introduced was within the framework of the International Computational Paralinguistics Challenge ComParE in 2016. Organizers of the challenge also presented a speech corpus that included deceptive and truthful speech samples, Deceptive Speech Database (DSD) [1], and a set of acoustic features based on the software tool openSMILE [5]. Base system proposed by organizers achieved results in terms of unweighted average recall (UAR) = 68.3%. The winners of the challenge [12] were able to achieve the UAR = 74.9%. They used prosodic features with base acoustic feature set. Later, in 2017 another system was proposed [11]. Authors used acoustic and lexical features with the classifier based on the model of Random Forest. They achieved the result in terms of F-score = 63.9% and Precision = 76.1%. In paper [20], authors proposed methods for the task of data scarcity and imbalanced classes in data for training models. In their system authors used SMOTE method for augmentation of training data with the number of k-nearest neighbours equal 3. This system was based on Support Vector Machines (SVM) and achieved results in term of UAR = 73.5%, mean F-score = 75.0% and Precision = 77.0%. In [21] the implementation of ensemble methods and neural networks were proposed. The ensemble consists of k-Nearest Neighbours, Random Forest and Neural Network have achieved the UAR of 65.0% and 70.0%, in case of average voting and majority voting correspondingly.

Pérez-Rosas et al. [13] proposed a multimodal corpus and a multimodal system for automatic detection of deceptive and truthful information. They used classifiers (Decision Trees and Random Forest) both for verbal and non-verbal features and achieved Accuracy in range of 60.0-75.0%. The same authors later presented another multimodal corpus that included deceptive and truthful samples [18]. The paper also proposed a model of Random Forest that achieved Accuracy = 69.0%. In [23] authors used corpora Box of Lies both for training and testing models. They extracted acoustic features from openSMILE, face markers and linguistic features for training models based on Random Forest. With this system they could achieve Accuracy of classification 73.0%.

Litvinova et al. [9, 10] used lexical and syntactic markers for deception detection in texts. They created a corpus [7] that consists of 226 essays on topic "Describe one day in your life". Every participant was free to answer truthfully or to lie. Besides, the corpus contains information about participants, namely: gender, age, scale of self-esteem, information collected using different psychological tests that can reveal the correlation between language parameters of written texts and personality characteristics of their authors. The mean length of texts is 221 word, all participants (46 men and 67 women) were native Russian speakers. Authors also applied statistic modelling with the use of Linguistic Inquiry and Word Count (LIWC). They proved the hypothesis that chosen marker of lie (relationship of percent of adverbs in text and percent of personal pronouns, thus, in truthful text percent of adverbs decreases and percent of personal pronouns increases) is effective. They achieved the detection of existence/absence of deceptive information with probability of 71.0-72.0%. In [8], after statistical analysis of created corpus authors found out that rates of Accuracy in classification differed for men and women – 73.3% and 63.3% respectively. In [14], authors used three groups of markers: psycholinguistic and sentiment markers, normalized frequencies of 11 Part-of-Speech (POS) tags and bigrams of POS tags, syntactic and readability features. They applied models of Random Forest and SVM. The best result was achieved with the use of SVM and POS tags and bigrams of POS tags. They also took into account the use of conjunctions, interjections and numerals. Such system achieved Accuracy = 57.0% and F-score = 56.0%.

2 Description of methods used in the present approach for automatic deception detection in speech

In order to process audio data, a researcher has to digitize and vectorize an audio signal. Modern automatic systems for paralinguistic analysis of speech use Low Level Descriptors (LLD) that represent spaces of feature vectors with a huge size (several thousands of features). These features are usually presented as feature sets (as in software toolkit openSMILE) and include different spectral, energy and prosodic features such as: fundamental frequency (F0), Mel-frequency cepstral coefficients (MFCC), formants or resonance frequencies of voice tract etc. In addition to those basic features, feature sets include their functionals: mean value, standard deviation, slope and shift, minimum value, relative position of minimum value, maximum value, relative position of maximum value, zero-crossing etc.

Boosting is a compositional machine learning meta-algorithm that is usually used for reduction of bias-variance tradeoff. Gradient boosting is a machine learning method that solves tasks of regression and classification using a composition of models for prediction. Boosting is a model with cascade process of training, where every following model attempts to correct the previous one. First, it creates a subset of data and the weights are equal for all training objects. Then the base model builds on this subset and makes a prediction for all set. After that it computes false predictions and updates weights (they gets bigger values). By this technique it builds other models for other subsets and makes predictions. Final model is the weighted mean of all models. Model of gradient boosting uses this technique of boosting and regression trees as a base algorithm, where every following tree builds on computed errors of previous one.

In the proposed approach we use three methods of gradient boosting: (1) Catboost, (2) XGBoost, (3) LightGBM. XGBoost (eXtreme Gradient Boosting) [19] is an implementation of the gradient boosting algorithm that has regularization function which helps to avoid the overfitting. LightGBM [6] is a faster implementation of gradient boosting and works especially good with big datasets. If compared with other implementations it builds trees in depth, not in breadth. Catboost [4] effectively copes with cat-

egorial variables and decreases time for their preprocessing, and it has a built-in detector of overfitting. Moreover, the approach uses the method of Stacking (Stacked Generalization) [3]. It is one of the ways to create an ensemble of algorithms. The idea is to combine output information of independent algorithms and use a classifier (or regressor) to make a final prediction.

3 Experimental setup

To solve the automatic detection of deceptive and truthful speech utterances task, we used two English-language corpora: (1) speech corpus Deceptive Speech Database and (2) multimodal corpus Real-Life Deception Detection Dataset (RLDDD) [13]. The first one consists of audio samples of students' speech. They played a role either a liar who stole papers from teacher's office, or an honest person. Recordings of the second corpus contain video data collected from public trial courts.

According to the other researches (for example, [22]), we decided to use both corpora simultaneously to increase the number of examples in training data, to get more generalized models and to improve the robustness of our models. Moreover, as it was found in [20], augmentation of data allows to improve results of models. In this work we use only audio features because we have only one corpus that contains video data. Overall number of audio recordings in both corpora is 1253 utterances. Small amount of data for training leads to significant restrictions in selection of machine learning methods and methods of modelling. This is also a reason of using augmentation of training data.

The proposed approach uses a software tool openSMILE to extract acoustic features. We chose three feature sets, namely: INTERSPEECH ComParE 2013 [15], ComParE 2016 (is an updated version of 2013 set) [5] and ComParE 2011 (includes acoustic features that were used in challenge for automatic detection of speaker state) [17]. Overall dimensionality of all sets was more than 12000 features. To balance classes in training data and perform an augmentation the method SMOTE (Synthetic Minority Oversampling Technique) was used. This method applies an algorithm of k-nearest neighbours that creates training objects similar to the minority class objects. Experimentally we found an optimal number of k-nearest neighbours equal 3. Due to the high dimensionality of a feature vector (more than 12000 features) there was a need to use a dimensionality reduction of feature space. To perform it the method of Principal Component Analysis (PCA) was applied. It is an implementation of programming library of Python language, Scikit-learn. As a result, we have got a set of 1680 training objects and 986 informative features. Right before the training process we shuffled the data thus each fold in 10-fold of cross-validation consists of data from both corpora and has similar distribution of classes.

To unite our models, we decided to use a two-level method of stacking, where the first level includes three models of gradient boosting and the second one contains a logistic regression. With the use of this method model on the second level makes predictions based on the results that it receives from the models on the first level. Overall scheme of such system including data preprocessing steps and two-level model for detection of deceptive and truthful information is shown on Fig.1.

For experiments we used methods of gradient boosting from three programming libraries: Catboost, XGBoost and LightGBM. Training and testing were performed with the use of 10-fold cross-validation to control and prevent overfitting. We also activated built-in overfitting detector in the Catboost. Hyper-parameters were selected empirically using grid search. As a quantitative rate F-score [2] and Unweighted Average Recall (UAR) were chosen.

4 Discussion of the results

The trained two-level model was able to achieve the quality of detection of deceptive and truthful information in speech in terms of F-score = 85.6%. For comparison, single models of Cabtoost, XGBoost and LightGBM have achieved results in terms of F-score of 84.1%, 84.6%, and 85.0% respectively. The achieved empirical results are highly competitive and comparable with the results presented by other researchers [11, 12, 17, 18, 21, 20, 23] (see Table 1). We present our results achieved with the use of two corpora and the approach described above. Moreover, the results show that the usage of several feature sets and models of gradient boosting can significantly improve the result in our task. However, as it was stated in [9] and other papers using several modalities can decrease variability in detection of

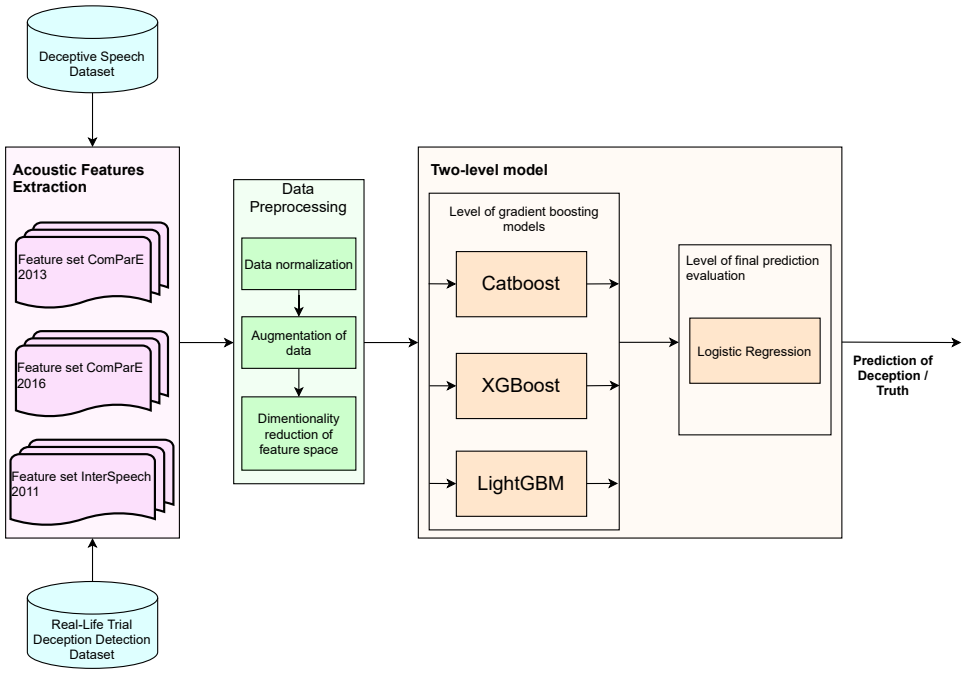

Figure 1: Architecture of the recognition system with two-level architecture for detection of deceptive and truthful information in speech.

deceptive and truthful speech utterances and improve results of the classification. We intend to check this hypothesis in future, namely we plan to analyze both lexical information and video data.

Approach	**Classification results**
Approach from [11]	F-score = 63.9%, Precision = 76.1%
Approach from [12]	UAR = 74.9%
Baseline system [17]	UAR = 68.3%
Approach from [18]	Accuracy (max) = 75.0%
Approach from [21]	UAR (max) = 70.0%
Approach from [20]	UAR = 73.5%, F-score = 75.0%, Precision = 77.0%
Approach from [23]	Accuracy = 73.0%
Catboost	F-score = 84.1%, UAR = 84.0%
XGBoost	F-score = 84.6%, UAR = 84.4%
LightGBM	F-score = 85.0%, UAR = 84.9%
Stacking (Catboost, XGBoost, LightGBM)	F-score = 85.6%, UAR = 85.5%

Table 1: Comparison of the achieved results with other known works.

5 Conclusions

The presented study is dedicated to the task of detection of deceptive and truthful information in speech that belongs to analysis of human's destructive behaviour, and this is a challenging task of computational paralinguistics. Existence of many scientific papers proves the significance of this task especially taking into account the widespread usage of the Internet and social networks. Due to restrictions in usage of contact-based methods contactless methods for deception detection in speech become more important. Since data retrieval for this task is time-consuming and difficult due to specificity of the task, the existing corpora have quite small amount of data and suffer from an imbalance in classes. To cope with these

restrictions, we have used an augmentation method (SMOTE) and a method for reducing feature space (PCA).

In the proposed approach, we have used a two-level model, where the first level includes three gradient boosting algorithms (Catboost, XGBoost, LightGBM) and the second one includes a logistic regression. The final prediction is based on predictions made on the first level. Hyper-parameters were calculated using the grid search method.

In the experiments, the proposed approach has achieved the quality of deception detection in terms of F-score = 85.6%. The proposed approach can be used to detect deceptive and truthful utterances, and gradient boosting methods can significantly improve results of the classification. The proposed approach can be applied as a component of a complex multimodal system for deception detection in speech with addition of analysis of lexical information and video data. It can also be a part of a prospective system for detection of human's psycho-physiological states and destructive behaviour.

Acknowledgements

This research was supported by the RFBR (project No. 20-37-90144), as well as by the Russian state research (No. 0073-2019-0005).

References

[1] B. Schuller. The INTERSPEECH 2016 computational paralinguistics challenge: deception, sincerity & native language. // Proceedings of INTERSPEECH-2016. — 2016. — P. 2001–2005.

[2] Chinchor N. MUC-4 Evaluation Metrics. // Proceedings of the Fourth Message Understanding Conference. — 1992. — P. 22–29.

[3] D. Wolpert. Stacked generalization. // Neural networks. — Vol. 5. — 1992. — P. 241–259.

[4] Dorogush A.-V. ErshovV. Gulin A. CatBoost: gradient boosting with categorical features support. // Workshop on ML Systems at NIPS 2017. — 2017.

[5] Eyben F. et al. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. // Proceedings of the 2013 ACM Multimedia (MM). — 2013. — P. 835–838.

[6] Ke G. Meng Q. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. // Advances in Neural Information Processing Systems. — 2017. — P. 3146–3154.

[7] Litvinova O. Litvinova T. et al. Deception detection in Russian texts. // Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics. — 2017. — P. 43–52.

[8] Litvinova T. A. Seredin P. V. et al. Text Classification on Basis of "False/True" using Methods of Automatic Text Processing. (In Russ.) // Nauchnyy dialog. — T. 10 (58). — 2016. — C. 70–83.

[9] Litvinova T.A. Litvinova O.A. A Study of Linguistic Features of Deceptive Texts with the Use of the Program Linguistic Inquiry and Word Count. (In Russ.) // Bulletin of the Moscow State Region University. Linguistics. — T. 4. — 2015. — C. 71–77.

[10] Litvinova T.A. Litvinova O.A. Text-Based Deception Detection: State-of-the-Art and Perspectives. (In Russ.) // Ivzestia of the Volgograd State Pedagogical University. Pedagogical sciences. — T. 2 (267). — 2015. — C. 189–192.

[11] Mendels G. Levitan S.I. et al. Hybrid acoustic-lexical deep learning approach for deception detection. // Proceedings of INTERSPEECH-2017. — 2017. — P. 1472–1476.

[12] Montaci´e C. Caraty M.-J. Prosodic cues and answer type detection for the deception subchallenge. // Proceedings of INTERSPEECH-2016. — 2016. — P. 2016–2020.

[13] Perez-Rosas V. Abouelenien M. et al. Deception Detection using Real-life Trial Data. // Proceedings of the ACM International Conference on Multimodal Interaction (ICMI 2015). — 2015. — P. 59–66.

[14] Pisarevskaya D. Litvinova T. Litvinova O. Deception Detection for the Russian Language: Lexical and Syntactical Parameters. // Proceedings of the 1st Workshop on Natural Language Processing and Information Retrieval associated with RANLP 2017. — 2017. — P. 1–10.

[15] R.K. Potapova. Variativnost' akusticheskih parametrov zvuchashhej rechi. [Variability of acoustic parameters of sounding speech]. (In Russ.) // Vestnik Moskovskogo gosudarstvennogo lingvisticheskogo universiteta. Gumanitarnye nauki. [Bulletin of Moscow State Linguistic University. Humanitarian sciences.]. — T. 740. — 2016. — C. 137–147.

[16] S.I. Smetanin. Toxic comments detection in Russian. // Proceedings of the International Conference Dialogue 2020. — 2020.

[17] Schuller B. Batliner A. et al. The INTERSPEECH 2011 speaker state challenge. // Proceedings of INTERSPEECH-2011. — 2011. — P. 3201–3204.

[18] Soldner F. P´erez-Rosas V. Mihalcea R. Box of Lies: Multimodal Deception Detection in Dialogues. // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — Vol. 1. — 2019. — P. 1768–1777.

[19] Tianqi Ch. Guestrin C. XGBoost: A Scalable Tree Boosting System. // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — 2016. — P. 785–794.

[20] Velichko A. Karpov A. Study of Data Scarcity Problem for Automatic Detection of Deceptive Speech Utterances. // Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019). — Vol. 2552. — CEUR-WS, 2020. — P. 38–46.

[21] Velichko A. Budkov V. et al. Applying Ensemble Learning Techniques and Neural Networks to Deceptive and Truthful Information Detection Task in the Flow of Speech. // Intelligent Distributed Computing XIII. IDC 2019. — Vol. 868. — Studies in Computational Intelligence, Springer, Cham, 2019. — P. 477–482.

[22] Velichko A.N. Budkov V.Yu. Karpov A.A. Study of classification methods for automatic truth and deception detection in speech [Issledovanie metodov klassifikatsii dlya avtomaticheskogo opredeleniya istinnoi ili lozhnoi informatsii v rechevykh soobshcheniyakh] // Science bulletin of the Novosibirsk state technical university [Nauchnyi vestnik Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta]. — 2018. — T. 3 (72). — C. 21–32.

[23] Zhang J. Levitan S.I. et al. Multimodal Deception Detection Using Automatically Extracted Acoustic, Visual, and Lexical Features. // Proceedings of Interspeech 2020. — 2020. — P. 359–363.