A Study of Data Scarcity Problem for Automatic Detection of Deceptive Speech Utterances*

Alena Velichko
alena.n.velichko@gmail.com

Alexey Karpov
karpov@iias.spb.su

St.Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS),
St.Petersburg, Russia

Abstract

In recent years there have been a lot of interest in the task of contactless automatic detection of deception in speech utterances. This task belongs to the paralinguistic area that studies the models for researching such aspects of speech as emotions, voice characteristics, psychophysiological traits etc. Despite the high relevance of the topic, there still remains the problem of the lack of data containing deceptive information. This paper presents an analysis of methods aimed to deal with the problem. Such over-sampling algorithms as SMOTE and ADASYN were explored.

Keywords: *Deception Detection in Speech, Computational Paralinguistics, Speech Technology, Machine Learning, Oversampling, SMOTE, ADASYN.*

1 Intoduction

Detection of deceptive and truthful speech utterances belongs to the area of computational paralinguistics as well as emotion recognition task, human psychophysiological states recognition task, addressee detection task, etc. All these tasks need a large amount of data to build a good classifier. Unfortunately, it is often quite difficult to collect a lot of data in current conditions. For example, existing databases used to learn models for deception detection consist of data where number of deceptive speech utterances is smaller than number of truthful speech utterances.

Thus, we have common problem of classification tasks – imbalanced data. It is a situation, when classes (we have two classes in our task – one for truthful and one for deceptive speech utterances) are presented with different number of objects. The problem of data scarcity is also encountered in such fields as fabricated news material detection, fake reviews for goods and products, telephone and online fraud, etc. In case of small difference in number of data, there is no need to do anything with imbalanced data. On the other hand, in case of big difference, we need at least change metrics for measuring classifiers performance to avoid "accuracy paradox" - a situation when accuracy of classifier is high but confusion matrix shows that the classifier always predicts majority class.

2 Related works

Although imbalanced data problem is common and severe in computational paralinguistic field, there is not so much works devoted to it. In [Schuller and Weninger, 2012] oversampling is mentioned as one of the recent trends in computational paralinguistic.

In [Chow and Louie, 2017] authors used Columbia-SRI-Colorado (CSC) corpus. They extracted acoustic, prosodic and lexical features and fed it to such models as logistic regression, support vector classifier and recurrent neural network for discovering lie correlated patterns. In the SMOTE experiments oversampling technique

was used along with lexical features and regularization. Results with oversampling were more optimistic than others. The best performance was achieved with the use of random sampling and bagging classifier in terms of accuracy of 58% and F1-score of 66%. The potential of oversampling technique was also presented in one of other paralinguistic tasks, depression detection [Cummins et al., 2014]. All experiments were conducted using Audio/Visual Emotion Challenge and Workshop (AVEC) 2013 dataset. This paper presented an oversampled extraction technique for the i-vector system in small datasets. Although oversampled classification accuracies of KL-means were lower than the standard ones, oversampled accuracies of i-vectors outperformed the standard ones by almost 16% in terms of mean accuracy. A study of imbalanced learning in sentiment analysis was presented in [Ah-Pine and Soriano-Morales, 2016]. Authors used three synthetic oversampling techniques (SMOTE, BorderLineSMOTE and ADASYN) for tweet-polarity classification. Oversampling improved the results of decision trees and l1 penalized logistic regression. Authors found out that all three methods improved recognition of the minority class as well as obtained large increase of the overall geometric mean criterion.

In [Kaya and Karpov, 2017] authors proposed another approach to handle the imbalanced data – weighted kernel classifiers. The pipeline combines suprasegmental acoustic features, FV encoding and multi-level normalization. The system effectively handled the class imbalance and did not need to use oversampling. The results were promising since they outperformed the baseline system results in Snoring Sub-Challenge and were on the same level as base results in the Addressee Sub-Challenge. Authors of [Akhtiamov et al., 2019] applied augmentation technique to addressee detection task in their cross-corpora experiments. They used an approach called mixup and it performed pretty good for neural networks with predefined acoustic features but did not gave a significant improvement in performance for e2e models, and did not benefit for linear classifiers and simple architectures without regularization at all. A novel imbalance learning-based framework for movie fear recognition was presented in [Zhang et al., 2018]. Authors conducted experiments using 4 different sampling techniques: SMOTE, Random Sampling, Hardsampling and Softsampling. The latter two methods were proposed to combine the advantages of oversampling and undersampling. The results reached a state-of-the-art performance on Recall and F1-measure in the MediaEval 2017 Emotion Impact of Movie Task. In [Ashihara et al., 2019] authors investigated whispered speech detection task and imbalanced learning impact on it. They used a class-aware sampling method for training phase and it helped to diminish the effect of imbalance in classes. The proposed system could achieve the best ROC-AUC score of almost 1.0 in close/neutral conditions and almost 0.9 in far-field condition.

The best result right now in deception detection in speech was achieved by [Mendels et al., 2017] and [Montacié et al., 2016] within the framework of ComParE-2017 and ComParE-2016 accordingly. Authors of the first paper presented a system that used acoustic and lexical features and Random Forest Classifier. The best performances in terms of F1-measure and precision were 63.9% and 76.1% accordingly. In the second paper authors elaborated a system with the use of prosodic ques, base feature set. The system reached the UAR of 74.9%. The base system [Schuller, 2016] on the competitions in 2016 performed with result in terms of UAR of 68.3%.

3 Technique Description

In case of imbalanced data, it is worthwhile to draw attention to the metrics for classifiers' performance evaluation. Such metrics as Precision, Unweighted Average Recall (UAR), F-score, Mean Squared Error (MSE) and other can help us to monitor the situation. Also, confusion matrix can be useful as well.

In case of binary classification task, confusion matrix in general looks like it is presented in Table 1. Formulas 1-5 present metrics we mentioned above. Precision focuses on False Positive errors while recall focuses on False Negative errors. UAR is mean of Recall of class 1 and Recall of class 2. F1-measure is a harmonic mean between Precision and Recall. MSE measures the average of the squares of the errors, in other words, the averaged squared difference between true values and predicted values.

There are two main ways of countering imbalanced data problem. In the first case, we need to remove a part of majority class objects, it is called undersampling. In the second case, we synthetically increase objects into minority class, it is called oversampling. It should be pointed out that it is better to use undersampling when overall number of data is more than tens and hundreds of thousands, and oversampling should be used when overall number of data is less.

$$Precision = \frac{TruePositive}{TruePositive + TrueNegative} \tag{1}$$

Table 1: General view of confusion matrix

	Class 1 predicted	**Class 2 predicted**
Class 1 Actual	True Positive	False Negative
Class 2 Actual	False Positive	False Negative

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{2}$$

$$UnweightedAverageRecall(UAR) = mean(Recall(Class1) + Recall(Class2)) \tag{3}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

$$MSE = \frac{\sum_{i-1}^{n} \times (y_i - \hat{y}_i)^2}{n} \tag{5}$$

There also exist other techniques of work with imbalanced data:

1. Collect more data;
2. Use other classifiers. For example, decision trees are good for imbalanced data classification (C4.5, C5.0, CART, Random Forest);
3. Use penalized models. Such models add weights to the model that makes wrong predictions to the minority class (penalized Linear Discriminant Analysis, penalized Support Vector Machines). This approach is appropriate when it is necessary to use certain classifier and when it is impossible to resample the dataset;
4. Use another concept such as anomaly detection and change detection.

Reducing number of objects in the majority class:

- Random undersampling. This technique implies counting number of objects in the majority class that should be removed to get optimal balance between classes. After that it randomly removes objects with or without replacement.
- Tomek links. Technique removes the majority class objects that overlap the minority class objects until all nearest neighbor pairs are of the same class. This method is widely used to remove noise from data [Tomek, 2010].
- Condensed Nearest Neighbor Rule. The main aim of using this method is to teach classifier to find differences between similar objects belonging to different classes [Hart, 1968].
- One-side sampling. This method combines two of the abovementioned methods. At first, it uses condensed nearest neighbor rule, then Tomek links method is applied [Kubat and Matwin, 1997].
- Neighborhood cleaning rule. The aim of this method is to remove all objects that affect adversely on the minority class objects classification. On the first step it classifies the data using 3-nearest neighbors method. On the second step it removes correctly classified majority class objects and neighbors of wrong classified majority class objects [Laurikkala, 2001].

Increasing number of objects in the minority class:

- Oversampling. Depending on the required balance between classes, this technique randomly chose minority class objects for copying or duplication.
- SMOTE (Synthetic Minority Oversampling Technique). As opposite to the abovementioned technique, SMOTE does not copy or duplicate minority class objects, it creates new similar objects. Using k-nearest neighbors method, SMOTE takes a vector between some minority class objects, then it multiplies the vector by random number between 0 and 1. New objects are created by summarization of obtained value and the initial value. Likelihood level of objects can be regularized by changing the parameter of k-nearest neighbors, also it is possible to set a number of objects to generate. Disadvantage of the method is increasing of minority class objects density that could affect on noise in data in case of equally distributed majority class objects [Chawla et al., 2002].
- ADASYN (Adaptive Synthetic Minority Oversampling). This technique was created similar to SMOTE but uses density function to automatically identify the number of objects that should be created for every minority class object. In so doing minority class objects weights change adaptively depending on the level of difficulty to learn classifier. Hence, new objects are creating mainly in front of minority class objects that are difficult to learn [Haibo et al., 2008].

In our case it was the most efficient way to change metrics for classifiers' performance evaluation, to use trees and oversampling. We chose the following metrics: Precision, Unweighted Average Recall, F-measure and Mean Squared Error. We decided to use oversampling of minority class, namely: ADASYN, SMOTE and two variants of SMOTE – BorderLineSMOTE (finds borderline minority class objects based on which new objects are creating) [Han et al., 2005] and SVMSMOTE (uses SVM method to identify minority class objects for creating new objects) [Nguyen et al., 2011].

4 Experiments

We used two databases to train and test models: Deceptive Speech Database [Schuller, 2016] and multimodal Real-Life Trial Deception Detection Dataset [Pérez-Rosas et al., 2015]. Following opensource toolkits were used: Ffmpeg to extract audio recordings from the multimodal dataset (https://www.ffmpeg.org/), Praat to preprocess audio data (http://www.fon.hum.uva.nl/praat/), openSMILE to extract 6373 low-level acoustic features (including pitch, energy, spectral and cepstral features, http://audeering.com/technology/opensmile/), Scikit-learn (to use implementations of the methods) [Pedregosa et al., 2011].

Then we oversampled each training set for every 10-cross validation, so testing sets did not contain synthetic data. Hence, training sets had balanced data. Total number of objects in training and testing sets were up to 1528 depending on the balance of classes in every training set.

Four classifiers were chosen according to the previous works [Velichko, Budkov et al., 2018; Velichko, Budkov et al., 2019]: Bagging with k-Nearest Neighbors as base classifier, k-Nearest Neighbors (k-NN), Support Vector Classifier (SVC) and Random Forest. Parameters of used methods were found with the use of grid search and presented in the Table 2. Figure 1 presents the proposed architecture of the system.

Table 2: Parameters of classifiers

Method	**Changed Parameters**
Bagging	base_estimator = KNeighborsClassifier (n_neighbors = 5, metric = 'manhattan', weights = 'uniform')
k-Nearest Neighbors	n_neighbors = 3, metric = 'manhattan', weights = 'uniform'
Random Forest	n_estimators = 300, min_samples_split = 5, min_samples_leaf = 1, bootstrap = False
Support Vector Classifier	C = 10, gamma = 0.0001, kernel = 'rbf'

The best results achieved with the use of above-mentioned classifiers and oversampling techniques are presented in Table 3. We used different number of nearest neighbors – 3 and 5 for all oversampling techniques.

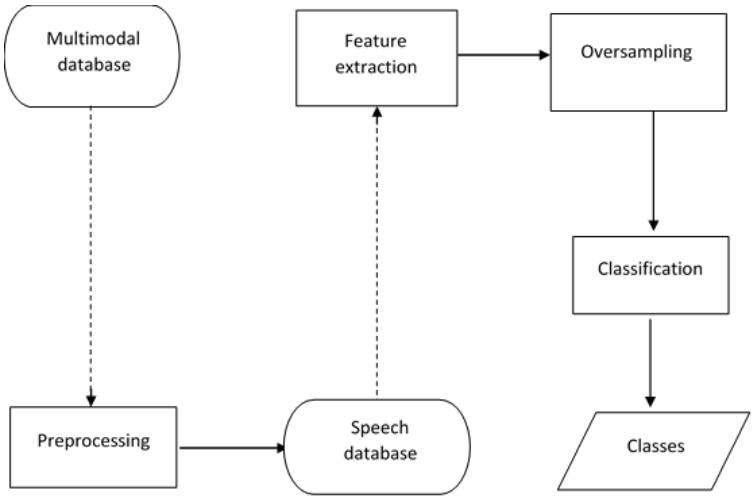

Figure 1: Overall architecture of the proposed system

Table 3: Best results achieved using four classifiers and difficult oversampling techniques

4*	**Bagging+k-NN**		**k-NN**		**Random Forest**		**SVC**	
	Precision / UAR / F1-score	**MSE**	**Precision / UAR / F1-score**	**MSE**	**Precision / UAR / F1-score**	**MSE**	**Precision / UAR / F1-score**	**MSE**
Without oversampling	**68.0** / **68.0** / **68.0**	0.49	**69.0** / **70.0** / **69.5**	**0.28**	**73.5** / 62.5 / 62.5	0.27	76.0 / 62.5 / 74.0	0.22
SMOTE (nn=3)	66.0 / 62.0 / 50.0	0.50	67.5 / 64.5 / 54.0	0.46	70.5 / 65.5 / 66.5	0.27	**77.0** / **73.5** / **75.0**	**0.21**
SMOTE (nn=5)	66.5 / 61.5 / 49.0	0.50	67.5 / 64.0 / 53.0	0.47	71.0 / 65.5 / 66.5	0.26	75.0 / 72.0 / 73.0	0.23
ADASYN (nn=3)	66.5 / 61.5 / 49.0	0.49	66.0 / 61.5 / 49.0	0.50	69.0 / 65.5 / 66.0	0.27	74.0 / 71.0 / 73.5	0.23
ADASYN (nn=5)	66.5 / 62.5 / 50.0	0.45	66.0 / 62.0 / 50.0	0.49	72.0 / 66.5 / 68.0	**0.25**	76.0 / 73.0 / 73.5	0.22
SVMSMOTE (nn=3)	66.5 / 64.5 / 54.5	**0.44**	68.5 / 66.0 / 56.0	0.44	71.5 / **67.5** / **68.5**	**0.25**	75.5 / 73.0 / 74.0	0.22
SVMSMOTE (nn=5)	66.5 / 64.5 / 55.5	**0.44**	66.0 / 64.5 / 54.0	0.46	71.0 / **67.5** / **68.5**	0.26	75.5 / 72.5 / 74.0	0.22
BorderLine SMOTE (nn=3)	65.5 / 62.0 / 51.0	0.49	68.0 / 64.0 / 52.0	0.47	69.5 / 63.5 / 64.5	0.27	74.5 / 72.5 / 73.0	0.23
BorderLine SMOTE (nn=5)	66.5 / 62.5 / 50.5	0.49	66.5 / 63.0 / 51.5	0.48	71.0 / 65.5 / 66.0	0.26	76.0 / 72.5 / 73.5	0.22

Table 4: Comparing results

2*System	Result		
	UAR, %	**Precision, %**	**F1-score, %**
Base system presented within ComParE-2016 [Schuller and Weninger, 2016]	68.3	-	-
System [Montacié et al., 2016]	**74.9**	-	-
System [Mendels et al., 2017]	-	76.1	63.9
Our previous system [Velichko et al., 2018]	71.0	-	-
Proposed system	73.5	**77.0**	**75.0**

5 Discussion

As we can see in tables above, the best results were achieved with the use of Support Vector Classifier and SMOTE oversampling technique with 3 neighbors. The results of the model were: UAR of 73.5%, mean F1-score of 75.0% and Precision of 77.0%.

The proposed deception detection system outperformed the results of base system presented in 2016 and our previous system [Velichko, Budkov and Karpov, 2018] in terms of UAR by 5.2% and 2.5% accordingly but underperformed the winners by 1.4% of UAR. Also, we outperformed result of the system presented in 2017 in terms of Precision and F1-score by 0.9% and 11.1% accordingly, see Table 4.

Results of experiments show that some of the methods proposed such as Random Forest and Support Vector Classifier can achieve good results with oversampled data and it is promising. Otherwise, such methods as Bagging and k-Nearest Neighbors did not achieve significant increase in performance with oversampled data, but worked quite good without sampling techniques.

6 Conclusions

In this paper we examined an important paralinguistic problem and deception detection task in particular, namely, data scarcity problem in classification task. Most popular techniques aimed to counter with imbalanced data were reviewed, the most suitable approaches for our task were chosen. A set of experiments was conducted with the use of 10-fold cross-validation method. Four classifiers with parameters found by grid search were used to find the best oversampling technique for our data. The best results were achieved using combination of the following methods: Support Vector Classifier with SMOTE oversampling technique, it resulted with 73.5% in terms of Unweighted Average Recall, mean Precision of 77.0% and mean F1-score of 75.0%. The proposed system can be used in such fields as banking area, prevention of telephone and online terrorism and fraud, in polygraph researches etc.

Acknowledgements

This research is supported by the Russian Science Foundation (project No. 18-11-00145).

References

[Schuller and Weninger, 2012] Schuller B., Weninger F.(2012) Ten Recent Trends in Computational Paralinguistics. *In: Esposito A., Esposito A.M., Vinciarelli A., Hoffmann R., Müller V.C. (eds) Cognitive Behavioural Systems. Lecture Notes in Computer Science*, vol 7403. Springer, Berlin, Heidelberg.

[Chow and Louie, 2017] Chow A., and Louie J. N.(2017) *Detecting Lies via Speech Patterns.*

[Cummins et al., 2014] Cummins, N., Epps, J., Sethu, V., Krajewski, J. (2014) *Variability compensation in small data: Oversampled extraction of i-vectors for the classification of depressed speech.* ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. 970-974. 10.1109/ICASSP.2014.6853741.

[Ah-Pine and Soriano-Morales, 2016] Ah-Pine, J., Soriano-Morales, E-P. *(2016) A Study of Synthetic Oversampling for Twitter Imbalanced Sentiment Analysis.* Workshop on Interactions between Data Mining and Natural Language Processing (DMNLP), Sep 2016, Riva del Garda, Italy. hal-01504684.

[Kaya and Karpov, 2017] Kaya, H., Karpov, A.A.(2017) *Introducing Weighted Kernel Classifiers for Handling Imbalanced Paralinguistic Corpora: Snoring, Addressee and Cold.* In: Proceedings of INTERSPEECH-2017, Stockholm, Sweden, pp. 3527-3531.

[Akhtiamov et al., 2019] Akhtiamov, O., Siegert, I., Karpov, A., Minker, W.(2019) *Cross-Corpus Data Augmentation for Acoustic Addressee Detection.* In: Proceedings of the SIGDial 2019 Conference, Stockholm, Sweden, pp. 274-283.

[Zhang et al., 2018] Zhang, X., Cheng, X., Xu, M., Fang Zheng, T.(2018) *Imbalance Learning-based Framework for Fear Recognition in the MediaEval Emotional Impact of Movies Task.* In: Proceedings of INTERSPEECH-2018, Hydebarad, India, pp. 3678-3682.

[Ashihara et al., 2019] Ashihara, T., Shinohara, Y., Sato, H., Moriya, T., Matsui, K., Fukutomi, T., Yamaguchi Y., Aono, Y.(2019) *Neural Whispered Speech Detection with Imbalanced Learning.* In: Proceedings of INTERSPEECH-2019, Graz, Austria, pp. 3352-3356.

[Mendels et al., 2017] Mendels, G., Levitan, S.I., Lee, K., Hirschberg, J.(2017) *Hybrid acoustic-lexical deep learning approach for deception detection.* In: Proceedings of INTERSPEECH-2017, Stockholm, Sweden, pp. 1472–1476.

[Montacié et al., 2016] Montacié, C., Caraty, M.-J.(2016) *Prosodic Cues and Answer Type Detection for the Deception Sub-Challenge.* In: Proceedings of INTERSPEECH-2016, San Francisco, USA, pp. 2016–2020.

[Schuller, 2016] Schuller, B.(2016) *The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity and Native Language* In: Proceedings of INTERSPEECH-2016, San Francis-co, USA, pp. 2001–2005.

[Tomek, 2010] Tomek, I.(2010) *Two modifications of CNN* In Systems, Man, and Cybernetics, IEEE Transactions on, vol. 6, pp 769-772.

[Hart, 1968] Hart, P.(1968) *The condensed nearest neighbor rule* In Information Theory, IEEE Transactions on, vol. 14(3), pp. 515-516.

[Kubat and Matwin, 1997] Kubat, M., Matwin, S.(1997) *Addressing the curse of imbalanced training sets: one-sided selection* In ICML, vol. 97, pp. 179-186.

[Laurikkala, 2001] Laurikkala, J.(2001) *Improving identification of difficult small classes by balancing class distribution.* Springer Berlin Heidelberg.

[Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P.(2002) *SMOTE: synthetic minority over-sampling technique* Journal of artificial intelligence research, 321-357.

[Haibo et al., 2008] He, H., Yang, B., Edwardo A. Garcia, and Shutao L.(2008) *ADASYN: Adaptive synthetic sampling approach for imbalanced learning* In IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322-1328.

[Han et al., 2005] Han, H., Wen-Yuan, W., Bing-Huan, M.(2005) *Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning* Advances in intelligent computing, 878-887.

[Nguyen et al., 2011] Nguyen, H. M., Cooper, E. W., Kamei, K.(2011) *Borderline over-sampling for imbalanced data classification* International Journal of Knowledge Engineering and Soft Data Paradigms, 3(1), pp.4-21.

[Pérez-Rosas et al., 2015] Pérez-Rosas V., Abouelenien M., Mihalcea R., Burzo M.(2015) *Deception detection using real-life trial data* Proceedings of ACM, pp. 59-66. doi:10.1145/2818346.2820758.

[Pedregosa et al., 2011] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E.(2011) *Scikit-learn: Machine Learning in Python* Journal of Machine Learning Research, vol. 12, pp. 2825-2830.

[Velichko, Budkov et al., 2018] Velichko A., Budkov V., Kagirov I., Karpov A.(2018) *Comparative Analysis of Classification Methods for Automatic Deception Detection in Speech* In Proc. 20th International Conference on Speech and Computer SPECOM-2018, Springer, LNAI vol. 11096, pp. 737-746.

[Velichko, Budkov et al., 2019] Velichko A., Budkov V., Kagirov I., Karpov A.(2020) *Applying Ensemble Learning Techniques and Neural Networks to Deceptive and Truthful Information Detection Task in the Flow of Speech* In: Kotenko I., Badica C., Desnitsky V., El Baz D., Ivanovic M. (eds) Intelligent Distributed Computing XIII. IDC 2019. Studies in Computational Intelligence, vol 868. Springer, Cham.

[Velichko et al., 2018] Velichko A.N., Budkov V.Yu., Karpov A.A.(2018) *Issledovanie metodov klassifikatsii dlya avtomaticheskogo opredeleniya istinnoi ili lozhnoi informatsii v rechevykh soobshcheniyakh [Study of classification methods for automatic truth and deception detection in speech]* Nauchnyi vestnik Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta – Science bulletin of the Novosibirsk state technical university, no. 3 (72), pp. 21–32. doi: 10.17212/1814-1196-2018-3-21-32 (In Rus.).