Effectiveness analysing of modification of training data set for voice forgery detection system

V E Mikhaltsov¹, Z V Semenova¹, a and E A Stepanova²

¹The Siberian State Automobile and Highway University, Mira pr. 5, Omsk, Russia
²Omsk Scientific-Research Institute of Instrument Engineering, 231, Maslennikova str., Omsk, Russia

Abstract. Paper presents data of the experiments with the advanced model of voice authentication. The main idea of the research is to prove or deny the hypothesis that adding generated fake data to the training set of the background model help to increase the probability of detecting fake data of the target announcer. Processing the results of a series of experiments did not confirm the hypothesis.

Keywords: voice authentication, deepfake, machine learning, training dataset, speaker model, generative adversarial network.

1. Introduction

Due to the high level of development of deep learning technologies, the problem of authenticity of voice and video data distributed on the Internet, in the media [1], used in digital forensics, etc., is especially acute. Machine learning algorithms can be used to determine the authenticity of digital media. Current publication activity in this field shows that systems for determining high-quality audio data forgeries (deepfake) created using various varieties of generative-adversarial neural networks (GANs) are of great interest to researchers [2].

Indicators of the volume of investments aimed at research in the field of machine learning indicate the acceleration of the development of speech synthesis technology [1]. At the same time, the global speech recognition and synthesis market is now one of the leading areas in the information technology industry.

The voice of any person is characterized by its acoustic features, such as tone, cadence, which are unique to each individual. To date, experts distinguish two main classes of biometric characteristics, according to which one can confirm a person's identity, these are behavioral and anatomical [3]. The main feature of voice biometry is that it is both an anatomical feature and behavioral. At the same time, as noted by A.K. Jain, A. Ross and S. Pancanti, in voice authentication and verification systems, the main problem is to checking the received reference values for compliance with those stored in the database [3].

Specialists distinguish verification models of two types [4]:

– content-dependent: text-dependent systems using a sample sound of a fixed phrase as a key-access to the system;
– text-independent: systems do not depend on the content and language in which the speech is delivered, and are built on the scenario of observing the voice carrier.
2. Problem Statement
The main idea is to test the hypothesis of engineers of the higher educational institution Zurich Kei Ishikawa, Jingqiu Ding and Xiaoran Chen, in which it was suggested that if you add the generated fake data to the training set of the background model, then the probability of detecting fake data of the target announcer will be higher than 50% [5]. These researchers used a background model trained only for natural voices. Thus, to test the hypothesis, it is necessary to prepare a background model containing a mixture of natural and synthesized voices in order to enable the neural network to learn to capture the features of the synthesized voice. Note that the published results do not provide detailed information on model measurements. The purpose of the experiments is to train background models and test their ability to detect voice spoofing attacks in accordance with this hypothesis.

3. Theory
First we need to prepare and configure environment, prepare datasets, clarify the training and test path, and select metrics. For the experiments, we used a Hyper-V virtualization environment and a PC with the following characteristics: 12-core AMD Ryzen 5 processor with a clock frequency of 3600 GHz; 40 GB of RAM; 128 GB SSD.

For the operation of the module, the OS Ubuntu 18.04 TLS is installed and Python 3.6 with following packets (according to the recommendations of the authors of the study [5]): googledrivedownloader (0.3); librosa (0.6.2); numpy (1.14.3); progressbar2 (3.37.1); pyworld (0.2.8); requests (2.20.0); scikit-learn (0.20.0); tensorflow (1.11.0); matplotlib (2.2.2).

It should be noted that the learning process requires a large amount of RAM. For continuous operation, it is recommended to increase the paging file and disable system protection against requests for allocation of large pages of memory. Protection is disabled by setting the following parameters in the boot area:

- vm.overcommit_memory = 2;
- vm.overcommit_ratio = 99.

Main module components. The adequacy of the experimental results largely depends on the understanding of the emerging space of feature sets in the correct way. As the number of private voice prints increases, the variety of features expands. This is due to the fact that the target voice of the speaker does not have exactly the same sound, therefore the neural network's knowledge of the features is not perfect.

We can identify the differences between the voices as follows:

1) analyze the features of synthesized voices and test the model's ability to detect differences in the sounding of a natural voice from a synthesized one;
2) analyze the features of voices close to the target and test the model's ability to distinguish almost identical voices.

4. Experiments
Let us present the results of a series of experiments that test the model's ability to detect differences in the sounding of a natural voice and a synthesized one. This module considers two possible scenarios in which there are:

- disjoint speaker model which knows some features from set A;
- shared speaker model, that has advanced knowledge of the disjoint model by additional analysis of the features of voice A, which the converter was trained.

The main module components are shown in table 1.

| No | Model type | Name       | Description                      |
|----|------------|------------|----------------------------------|
| 1. | Speaker    | GMM-disjoint | Knows speaker                    |
| 2. | Speaker    | GMM-shared | Knows speaker and initial voice   |
| 3. | Background | UBG        | Knows different voices            |
| 4. | Converter  | CycleGAN   | Knows initial voice and target speaker’s voice |
The models presented above were pre-trained and provided by the developer along with the source code, training and testing sets.

Training and testing datasets.

The target voice was 7 videos of Barack Obama's speeches, 15 minutes long, from which the following datasets were compiled:
- 3 videos for the speaker GMM model training;
- 3 videos for testing;
- 1 video to teach the converter CycleGAN Obama voice and GMM-shared.

Also, the module contains 300 minutes to train the UBG background model for the voices of different people.

To assess the quality of voice recognition, the authors [6] provided sound options for about 100 minutes of a high-quality fake. We need to prepare kits for training new background models. To make sure that the author's hypothesis is confirmed, it is necessary to identify the development of the model's ability to capture the features of the synthesizer in the sound of the voice. For this, training sets were compiled with a variable ratio of natural to synthesized voices. The background model provided by the researchers contains 2000 audio files with a total sound of 300 minutes. We used the following natural versus synthesized file ratios: 2000 : 0 (GMM-UBG); 1750 : 250 (GMM-UBG 250); 1500 : 500 (GMM-UBG 500); 1250 : 750 (GMM-UBG 750); 1000 : 1000 (GMM-UBG 1000).

According to the technological requirements for working with a private GMM-UBG model, you should use voices belonging to the same gender. Failure to comply with this condition leads to distortion of the results.

The voices are generated by various algorithms and are taken from an open set provided by ASV Spoof 2020. It took 24 hours to train background models with different ratios. All test suites are provided by the developer and are used without modifications to evaluate the effectiveness.

The symbols for predictions resulting from testing in accordance with the set used are defined in Table 2.

| Graphs symbol       | Appropriate dataset                                           |
|---------------------|----------------------------------------------------------------|
| Test                | Testing dataset for the speaker model                         |
| Fake                | Obama’s fake voice                                             |
| Universal background| Testing dataset for the background model                       |
| Train_conv          | Training dataset for the converter                             |
| Validation verification | Training dataset for the speaker model                        |

Source code «train_and_plot.py» contains python instructions for the training, testing and prediction results saving. It is taking enough time to training neural models, so we have excluded the function of collecting predictions to speed up the program. As a result of functional changes, 4 models are trained continuously.

**Metrics selection.**

Senior Machine Learning Engineer and Editor of NeptunAI Journal J. Czakon, in his review, describes metrics for evaluating the performance of neural networks and notes that using most of them implies that you have balanced data for analysis [7]. Balanced data is the set with equal number of positive and negative samples. In a real situation, it is not always easy to ensure the balance of data, therefore, to analyze the behavior of the classifier at different threshold values, the most often used graphs of the operating characteristic (ROC-curve, Receiver Operating Characteristics curve), which shows accurate results for balanced data and a curve that is sensitive to the ratio of classes (PR-curve, Precision-recall-curve) for unbalanced. Both ROC and PR curves are based on predicted estimates of classification models. Their difference lies in the fact that the ROC curve is based on the number of true positive (TPR, true positive rate) and false positive results (FPR, false positive rate), while the PR curve is based on the correctness of the predicted value (PPV, positive predict value) and TPR [6].
In our case dataset is poorly balanced: 600 positive (natural target voice samples) and 400 negative (fake target voice samples). Therefore, for the analysis as a metrics of assessing GMM-UBG classifiers, it was decided to use comparative indicators: PR-curve; ROC curve; calculations of the area under the ROC and PR curve showing quantitative estimates of the quality of the models.

The developer provided the code for obtaining the logarithmic probabilities of the models in the "train_and_plot.py" file, which only worked with 3 models. We have 7 models at our disposal after training new models. For the convenience of working with 7 models in one run, as well as sorting the obtained forecasts, the author's source code has been modified.

The author's code, described in the file "compute_auc.py", needed to be modified for the following reasons. First, some metrics need to be implemented, namely the following have been added: the "roc_curve ()" function for drawing the ROC curve; function "precision_recall_curve ()" for drawing the PR curve; function "average_precision_recall ()" to get the area under the PR curve. Secondly, it is necessary to display the results obtained on graphs. Thirdly, for the convenience of subsequent work with estimates, the results are sorted into directories, corresponding to the name of the background model.

Thus, as a result of the preparatory work, a working environment was set up, assessment metrics were selected, the corresponding code in Python was developed, and background models were prepared for experimental research.

5. Results
At the first stage of the experimental work, we prepared and reproduced the experiment of the authors of the hypothesis and obtained results that coincide with the results of their experiments, namely, the repeated analysis of the author's model GMM-UBG confirms a good difference between voices from the universal background and Obama's voice.

The authors proposed to train the background model for synthesized voices. Figure 1 shows histograms for the created models in accordance with the hypothesis of the researchers.

Figure 1. Histograms for trained disjoint models

Figure 1 presents that proportional changes in the sound of natural and synthesized voices do not lead to an improvement in their diarization, that is, to the differentiation of voices.
Table 3 presents the average estimates of the distributions obtained as a result of the forecast of each of the models, which, in fact, are the centers of the obtained classes.

### Table 3. Average logarithmic forecast for disjoint models

| Data set                  | GMM-UBG 250 | GMM-UBG 500 | GMM-UBG 750 | GMM-UBG 1000 |
|---------------------------|-------------|-------------|-------------|--------------|
| Test                      | 1.08        | 1.1         | 1.14        | 1.28         | 1.4          |
| Universal background      | -9.09       | -9.0        | -8.89       | -8.74        | -8.54        |
| Fake                      | 1.2         | 1.3         | 1.44        | 1.58         | 1.61         |
| Training (for converter)  | 1.24        | 1.32        | 1.36        | 1.45         | 1.53         |
| Validation verification   | 5.22        | 5.3         | 5.44        | 5.56         | 5.69         |

### Table 4. Average logarithmic forecast for shared models

| Data set                  | GMM-UBG 250 | GMM-UBG 500 | GMM-UBG 750 | GMM-UBG 1000 |
|---------------------------|-------------|-------------|-------------|--------------|
| Obama's voice             | 1.89        | 1.88        | 1.92        | 2.06         | 2.18         |
| Other people's voices     | -8.93       | -8.84       | -8.73       | -8.58        | -8.37        |
| Obama's fake voice        | 1.56        | 1.67        | 1.8         | 1.94         | 1.98         |
| Obama's voice (A)*        | 3.9         | 3.98        | 4.02        | 4.11         | 4.19         |
| Obama's voice (B)**       | 5.2         | 5.29        | 5.42        | 5.54         | 5.67         |

(A)* - Obama's voice which used for training partially of speaker model and the converter, (B)** - Obama's voice which used for training of speaker model.

Analysis of the data in Tables 3 and 4 indicates that when the proportion changes towards increasing the synthesized sound, there is a general shift in the positive direction. In order to correctly interpret the obtained estimates, it is necessary to convert the predicted logarithmic probability into a binary form. Average scores are presented in tables 5 and 6.

### Table 5. Average binary forecast for disjoint models

| Data set                  | GMM-UBG 250 | GMM-UBG 500 | GMM-UBG 750 | GMM-UBG 1000 |
|---------------------------|-------------|-------------|-------------|--------------|
| Obama's voice             | 0.75        | 0.75        | 0.76        | 0.78         | 0.80         |
| Other people's voices     | 0.00        | 0.00        | 0.00        | 0.00         | 0.00         |
| Obama's fake voice        | 0.77        | 0.79        | 0.81        | 0.83         | 0.83         |
| Obama's voice used to train the converter | 0.78 | 0.79 | 0.80 | 0.81 | 0.82 |
| Obama's voice used to train the speaker model | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 |

Analyzing the data from Table 5, we can conclude that there is an improvement in the quality of natural voice recognition. Moreover, the verification accuracy is 99%. However, in the situation with synthesized voices, the models still do not distinguish between them. At the same time, confidence in the truth of the synthetic voice even began to somewhat prevail in relation to the natural sound.

Analysis of the results presented in Table 6 for shared models also show an improvement in test voice recognition. Note that, as fake data is added to the training sample of the background model, it can be seen that the difference between the real voice of the speaker and the synthesized one began to decrease. This shows that, in general, the decision to modify the model led to a deterioration in the diarization of the synthetic voice from the real voice of the speaker. In order to make an accurate conclusion about the results obtained, it is necessary to construct ROC and PR curves. ROC curves for disjoint and shared models are shown in Figure 2.
Table 6. Average binary forecast for shared models

| Data set                                      | GMM-UBG | GMM-UBG 250 | GMM-UBG 500 | GMM-UBG 750 | GMM-UBG 1000 |
|-----------------------------------------------|---------|-------------|-------------|-------------|--------------|
| Obama's voice                                | 0.87    | 0.87        | 0.87        | 0.89        | 0.90         |
| Other people's voices                        | 0.00    | 0.00        | 0.00        | 0.00        | 0.00         |
| Obama's fake voice                           | 0.83    | 0.84        | 0.86        | 0.87        | 0.88         |
| Obama's voice used to train the converter and partially speaker model | 0.98    | 0.98        | 0.98        | 0.98        | 0.99         |
| Obama's voice used to train the speaker model | 0.99    | 0.99        | 1.00        | 1.00        | 1.00         |

Figure 2. ROC curves for disjoint and shared models

The bottom left point (0; 0) represents a strategy in which the classifier never gives false positives, but also does not give true positives. Point (1; 1) reflects the opposite situation. The ideal classifier is the one that passes through the point (0; 1). Thus, the classifier that shows the bend in the upper left region and is closest to the point (0; 1) is considered more accurate. [7].

The diagonal line on the graphs, where y = x presents an example of random classification. As is known, random classifiers show a sliding graph along a diagonal line [7].

The lower the bend in the lower right region, the worse the classifier than the random one. This indicates that he begins to make negative decisions for the truth. At the same time, it is believed that the model that gives a bend below the diagonal line has sufficient knowledge about the speaker, but does not know how to use them correctly [7].

In order to determine exactly which classifier works more correctly, the areas under the ROC curve are calculated (table 7) [7].

Table 7. The estimate of area under an ROC curve for disjoint and shared models

| Model          | For disjoint model | For shared model |
|----------------|--------------------|------------------|
| GMM-UBG        | 0.53               | 0.61             |
| GMM-UBG 250    | 0.51               | 0.60             |
| GMM-UBG 500    | 0.50               | 0.58             |
| GMM-UBG 750    | 0.50               | 0.58             |
As a result of the analysis data in Table 7, we can say that the models trained only in natural voice show results slightly higher than those trained in mixed sound with synthesized voices. At the same time, indicators below 0.75 are considered unsatisfactory.

Precision-recall curves. J. Davis and Professor M. Goadrich in their research [8] prove that the points of the ROC-curve have a one-to-one comparison between the points of the PR-curve. Thus, according to the theorem, where for a fixed number of positive and negative patterns, one curve dominates the other in the ROC space if and only if the first one dominates the second in the PR space. This approach provides to make an optimal choice between the models, unambiguously determining which classifier copes with the task in the best way. The PR curves show positive predictive scores on the ordinate and true positives on the abscissas (Figure 3). Moreover, the model whose PR-curve is closest to the upper right corner is considered the best [8].

![Average precision-recall for disjoint models](image)

Figure 3. PR curves for disjoint and shared models

In Figure 3, you can see the accuracy versus completeness. At the same time, the classifier that covers a large area under the curve is considered stronger. The curves start with 100% accuracy, with TPR = 0.5 position, since PPV is not determined when the denominator (TP + FP) is zero, which means that no false positives have occurred. Then, a decrease in the curve is observed almost strictly downward, showing a drop in accuracy. At the same time, the number of true positive positives decreases, which indicates an increase in false positives. Further, the curve diagonally descends uniformly to the point (1; 0), demonstrating the same number of correct positive predictions with false positives, which indicates that the models have little idea of the difference between the two classes. However, the shared model retains prediction accuracy a little bit longer.

Table 8 shows the estimated area of the PR-curve for disjoint and shared models, respectively.

| Model       | PR-scorefordisjoint | PR-scoreforshared |
|-------------|---------------------|-------------------|
| GMM-UBG     | 0.65                | 0.72              |
| GMM-UBG 250 | 0.64                | 0.71              |
| GMM-UBG 500 | 0.63                | 0.70              |
| GMM-UBG 750 | 0.63                | 0.69              |
| GMM-UBG 1000| 0.64                | 0.71              |
Estimated areas under the ROC and PR curves for disjoint and shared models allowed us to perform a comparative analysis. The analysis showed that changing the proportion of the training set of the background model towards synthesized samples does not lead to an increase in the forecast accuracy. Despite the fact that the shared model has higher scores, it is taken to consider a particular case, which may not occur in real practice. Thus, the estimates obtained by the disjoint models show that the situation of random classification is preserved. The results of the conducted research have shown the inability of the models to distinguish synthesized sounding of voices from natural ones.

6. Conclusion

Results shows that the hypothesis that if you add the generated fake data to the training set of the background model, then the probability of detecting fake data of the target speaker will become higher than 50% was not confirmed.

A possible solution would be to change the MFCC method for extracting features from a voice with another more modern one. It is worth trying to go through the analysis of fake voices that have an exceptional similarity with the target speaker, adding to the background model of fake this target speaker, generated by various algorithms. And also it is worth investigating the possibility of the work of the discriminator of the CycleGAN model to help in identifying features, through the use of the knowledge it gained about the target speaker.

Also, the disadvantage of the presented approach is that Obama’s voice is specific in overtones that it differs significantly from all others. Alternative metrics for evaluating the results should be considered.

References

[1] Kaliyar R 2021 DeepFak E: improving fake news detection using tensor decomposition-based deep neural network The Journal of Supercomputing 10.1007/s11227-020-03294-y
[2] Jain A K, Ross A, Pankanti S 2006 Biometrics: a tool for information security IEEE Trans. Inf. Foren. Secur. 1 pp 125–143
[3] Thanh Thi Nguyen 2020 Deep Learning for Deepfakes Creation and Detection: A Survey Available at: https://arxiv.org/abs/1909.11573 p 1–10
[4] Alegre F, Evans, Kinnunen T, Wu Z 2014 Anti-Spoofing: Voice Database (Springer US) p 1–10
[5] Ishikawa K, Ding J, Chen X 2018 Can we detect fake voice generated by GANs? Available at: https://github.com/kstoneriv3/Fake-Voice-Detection/blob/master/DL.project_fake_voice_detection.pdf (assessed on: 25.06.2021)
[6] Czakon J 2019 Ultimate guide to binary classification metrics Available at: https://towardsdatascience.com/the-ultimate-guide-to-binary-classification-metrics-c25c3627dd0a#cd01;
[7] Fawcett T 2004 ROC Graphs: Notes and practical considerations for researchers in Pattern Recognition Letters vol. 31 2004p 1–38
[8] Davis J, Goadrich M 2006 The relationship between Precision-Recall and ROC curves in Proceedins of the 23rd International Conference on Machine Learning (Pittsburg, PA) pp 1–8