УНИВЕРСИТЕТ ИТМО
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2021 Том 21 № 1 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2021 Vol. 21 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2021-21-1-109-117

Methods of countering speech synthesis attacks on voice biometric systems in banking (review article)

Aleksandr Yu. Kuznetsov[1], Roman A. Murtazin[2], Ilnur M. Garipov[3], Evgeny A. Fedorov[4], Anna V. Kholodenina[5], Alisa A. Vorobeva[6]✉

[1,2,3,4,5,6] ITMO University, Saint Petersburg, 197101, Russian Federation
[4] Laboratory PPS Ltd, Saint Petersburg, 199178, Russian Federation

[1] al.ur.kouznetsov@gmail.com, https://orcid.org/0000-0002-5702-3786
[2] murtazinroman3161@gmail.com, https://orcid.org/0000-0003-3669-7586
[3] i_garipov@mail.ru, https://orcid.org/0000-0003-3108-5484
[4] fyodorov.1997@gmail.com, https://orcid.org/0000-0003-2911-5509
[5] annaholodenina@gmail.com, https://orcid.org/0000-0003-1911-3710
[6] vorobeva@itmo.ru✉, https://orcid.org/0000-0001-6691-6167

Abstract

The paper considers methods of countering speech synthesis attacks on voice biometric systems in banking. Voice biometrics security is a large-scale problem significantly raised over the past few years. Automatic speaker verification systems (ASV) are vulnerable to various types of spoofing attacks: impersonation, replay attacks, voice conversion, and speech synthesis attacks. Speech synthesis attacks are the most dangerous as the technologies of speech synthesis are developing rapidly (GAN, Unit selection, RNN, etc.). Anti-spoofing approaches can be based on searching for phase and tone frequency anomalies appearing during speech synthesis and on a preliminary knowledge of the acoustic differences of specific speech synthesizers. ASV security remains an unsolved problem, because there is no universal solution that does not depend on the speech synthesis methods used by the attacker. In this paper, we provide the analysis of existing speech synthesis technologies and the most promising attacks detection methods for banking and financial organizations. Identification features should include emotional state and cepstral characteristics of voice. It is necessary to adjust the user's voiceprint regularly. Analyzed signal should not be too smooth and containing unnatural noises or sharp interruptions changes in the signal level. Analysis of speech intelligibility and semantics are also important. Dynamic passwords database should contain words that are difficult to synthesize and pronounce. The proposed approach could be used for design and development of authentication systems for banking and financial organizations resistant to speech synthesis attacks.

Keywords

biometrics, automatic speaker verification, banking authentication, synthetic speech, spoofing detection

Acknowledgements

The paper was prepared at ITMO University within the framework of the scientific project No. 50449 "Development of cyberspace protection algorithms for solving applied problems of ensuring cybersecurity of banking organizations".

For citation: Kuznetsov A.Yu., Murtazin R.A., Garipov I.M., Fedorov E.A., Kholodenina A.V., Vorobeva A.A. Methods of countering speech synthesis attacks on voice biometric systems in banking. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2021, vol. 21, no. 1, pp. 109–117. doi: 10.17586/2226-1494-2021-21-1-109-117

УДК 004.934.8’1; 004.056.53

Методы противодействия атакам посредством синтеза речи на голосовые биометрические системы в банковской сфере (обзорная статья)

Александр Юрьевич Кузнецов[1], Роман Андреевич Муртазин[2], Ильнур Мидхатович Гарипов[3], Евгений Андреевич Фёдоров[4], Анна Викторовна Холоденина[5], Алиса Андреевна Воробьева[6]✉

[1,2,3,4,5,6] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[4] ООО «Лаборатория ППШ», Санкт-Петербург, 199178, Российская Федерация

[1] al.ur.kouznetsov@gmail.com, https://orcid.org/0000-0002-5702-3786
[2] murtazinroman3161@gmail.com, https://orcid.org/0000-0003-3669-7586
[3] i_garipov@mail.ru, https://orcid.org/0000-0003-3108-5484
[4] fyodorov.1997@gmail.com, https://orcid.org/0000-0003-2911-5509
[5] annaholodenina@gmail.com, https://orcid.org/0000-0003-1911-3710
[6] vorobeva@itmo.ru✉, https://orcid.org/0000-0001-6691-6167

Аннотация

Рассмотрены методы противодействия атакам синтеза речи на банковские голосовые биометрические системы. Безопасность голосовых биометрических систем является масштабной проблемой, значительно развивающаяся в последние годы. Системы автоматической верификации говорящего (ASV) уязвимы для различных типов спуфинг-атак: имперсонализация, повторное воспроизведение, преобразование и синтез речи. Технологии синтеза речи стремительно развиваются (GAN, Unit selection, RNN и др.), поэтому такие атаки сегодня наиболее опасны. Показано, что противодействие спуфинг-атакам может быть основано на поиске аномалий фазы и частоты тона, которые появляются во время синтеза речи, а также на предварительном знании акустических различий конкретных синтезаторов речи. Безопасность ASV остается нерешенной проблемой, не существует универсального решения, которое бы не зависело от используемых злоумышленником методов синтеза речи. Представлен анализ существующих технологий синтеза речи. Рассмотрены наиболее перспективные методы обнаружения атак для банковских и финансовых организаций. Комплекс мер должен учитывать эмоциональное состояние клиента банка, кепстральные характеристики голоса. Необходима регулярная корректировка голосового отпечатка пользователя для поддержания его актуальности. Анализируемый сигнал не должен быть слишком плавным, содержать неестественные шумы, резкие перерывы, изменения уровня сигнала. Важное значение имеют внятность речи, выявление и учет ее семантических особенностей. База динамических паролей должна содержать сложно синтезируемые и произносимые слова. Предлагаемый подход может быть использован для проектирования и разработки систем аутентификации для банковских и финансовых организаций, устойчивых к атакам синтеза речи.

Ключевые слова

биометрия, распознавание по голосу, аутентификация в банковской сфере, синтезированная речь, выявление фальсификации голоса

Благодарности

Работа выполнена в Университете ИТМО в рамках темы НИР № 50449 «Разработка алгоритмов защиты киберпространства для решения прикладных задач обеспечения кибербезопасности организаций банковской сферы»

Ссылка для цитирования: Кузнецов А.Ю., Муртазин Р.А., Гарипов И.М., Фёдоров Е.А., Холоденина А.В., Воробьева А.А. Методы противодействия атакам посредством синтеза речи на голосовые биометрические системы в банковской сфере // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21, № 1. С. 109–117 (на англ. яз.). doi: 10.17586/2226-1494-2021-21-1-109-117

Introduction

The advent of the technologies that allow accurately identifying persons by their biometric characteristics and the development of image recognition technologies have stimulated the improvement of biometric identification systems. The future of biometrics has attracted much attention from researchers, creating a wide variety of modalities currently used by biometric systems. Following the researchers, many organizations from various fields, including banking, began to implement and use biometrics for authentication, identification, access control, and other purposes.

Today, most identification/authentication systems are based on analyzing biometric samples publicly available for observation, such as fingerprints, face and palm geometry, voice, and some others. The widespread use of these modalities is due to the availability of the necessary reading equipment and low implementation costs. All of these modalities make it possible to authenticate or identify subjects, but they differ in the value of type I and type II errors: FRR (False Rejection Rate) and FAR (False Acceptance Rate).

Static biometric samples, which are robust to changes over time, have sufficient uniqueness and allow high accuracy of person recognition. However, they are more vulnerable to attacks from malefactors than dynamic biometric samples, as they can change at the will of the subject (as is the case, for example, for voices or handwritten signatures). Static biometric samples can

be more easily falsified. For example, there are several technologies for creating fake fingerprints [1], and 3D masks and even photographs can be used for such purposes.

Automatic Speaker Verification (ASV) systems are a promising area for improving identification quality. The collection of voice data does not require a special capture device, as it can be obtained using a standard sound recorder available in smartphones, stationary phones, tablets, or laptops. Voice is a dynamic characteristic that increases the reliability of biometric systems and creates a great identification potential for their use.

The use of ASV for remote identification/authentication can allow banks to extend the availability of their financial services, such as opening an account or obtaining a loan, while freeing customers from having to visit bank branches in person.

According to J'son & Partners Consulting, an analytics company, the average annual growth rate of the ASV market by 2022 is estimated at 21.12 %. It is higher than the growth rate of systems with other biometric modalities.

A number of independent studies have confirmed the vulnerability of ASV technology to spoofing threats. However, compared to verification involving other biometric modalities, spoofing and countermeasure research for ASV is still in its infancy. A current barrier to progress is a lack of standardization, making it difficult to compare results generated by different researchers. The ASVspoof initiative aims to overcome this through the provision of standard corpora, protocols, and metrics to support common evaluations.

Attackers can use ASV spoofing techniques for various types of fraud, including identity theft. For the banking sphere, this issue is the most relevant one due to the amount of information and services that users gain access to upon successful authentication. Even rare cases of information leakage can lead to significant financial and reputational losses.

Known ASV vulnerabilities include spoofing attacks through impersonation, replay, voice conversion, and speech synthesis. The latter type of attack is the most dangerous.

The relevance of this research is due to

1) an increase in the number of ASV spoofing technologies based on voice synthesis;
2) the development of technologies for falsifying biometric characteristics, such as DeepFake voice, which is based on Generative Adversarial Networks (GANs);
3) the imperfection of methods used to counter attacks.

It is impossible to use remote voice biometric authentication systems without implementing methods of counteracting spoofing attacks. Therefore, there is a need to develop a means of detecting speech synthesis attacks, which would increase the security of ASV.

The architecture of a Typical Automatic Speaker Verification System

Biometric identification/authentication systems rely on the individual biometric characteristics of a person and perform the function of automatic identity recognition. Unlike other systems, ASV uses both physiological (static) and behavioral (dynamic) user features (voice and speech) [2].

ASV works similarly to other identification systems, such as those based on fingerprints or faces. However, it is necessary to reduce noise and areas which do not contain useful signal components for required audio signal preprocessing.

To ensure the correct operation of ASV for identification and verification, users must be registered in a database. Figure 1 shows the ASV model architecture as a process that includes the stages of data input, signal pre-processing, feature extraction, template creation (registration), and comparison with a template (verification, identification) [3].

The second stage is usually performed jointly with the third stage. At the stage of comparison with the template, the system returns the results of searching users closest to the presented voice sample from the database when identifying the subject. At the same stage, when verifying the subject, the system returns the probability of the subject's image coinciding with the template.

Many researchers propose using cepstral coefficients as features. To calculate the cepstrum, the signal is passed

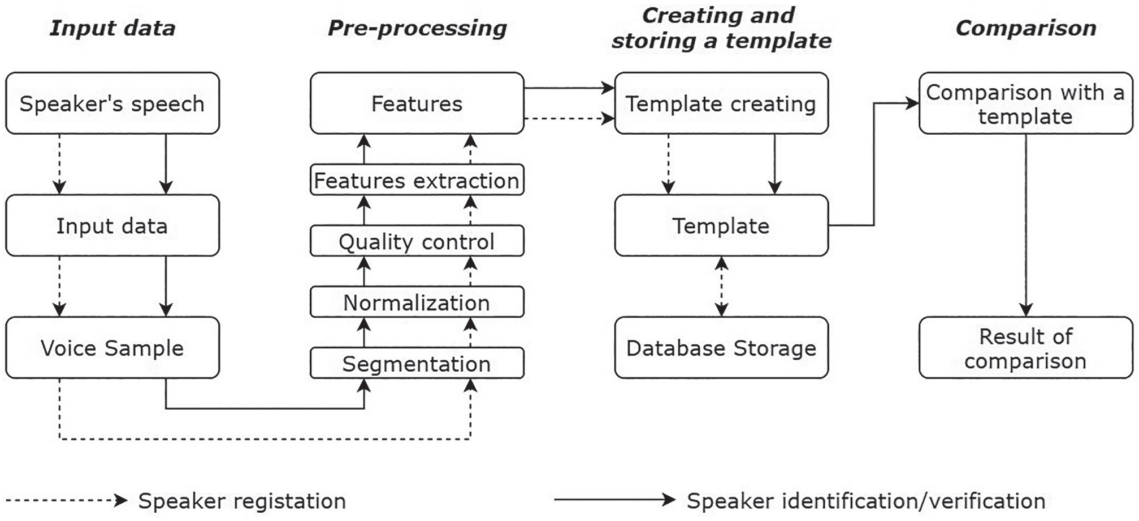

Fig. 1. The ASV architecture

through a high-frequency amplifying filter, divided into equal frames. Then, the frequency spectrum is smoothed at the frame boundaries using the window function. Afterward, the spectrum is found and multiplied with the spectrum of the received filter bank using the Fourier transform. The logarithm is taken from the resulting spectrum envelope and the inverse discrete Fourier transform is applied.

$$c[n] = F^{-1}\{\ln|F\{x(n)\}|\},$$

F — forward Fourier transform, F^{-1} — reverse Fourier transform.

Mel-frequency cepstral coefficients are most commonly used to describe the characteristics of a phoneme [4]. To convert the frequency from Hz to Mel, the following formula is used:

$$Mel(f) = 2595\lg(1 + \frac{f}{700}),$$

f — frequency, Hz.

The Hamming window is the most popular choice among other window functions such as the rectangular window, Kaiser window, Blackman window, and Hann window. Cochlear, linear frequency, bark frequency, and other cepstral coefficients are also used in research. In practice, other features make it possible to form a unique speaker template, which helps to distinguish the amplitude-frequency, spectral, spectral-temporal, formant, and phase characteristics. There are also methods for automatic speaker recognition through calculating the transfer function of the vocal tract, such as Joint Factor Analysis (JFA), Total Variability Matrix (TVM), and Probabilistic Linear Discriminant Analysis (PLDA).

Also, there are a large number of methods and algorithms for synthetic speech detection, including Gaussian Mixture Models (GMM) [5], Support Vector Machines (SVM) [6], Neural Networks (NN) [7], Hidden Markov Models (HMM) [8], Relative Phase Shift (RPS) [9], and Vector Quantization (VQ) [10].

The variability of building voice authentication systems is increasing. The lack of an effective combination of extracted speech features, input signal processing methods, and auxiliary functions complicates the search for the optimal algorithm to implement ASV in the banking sphere.

Main Types of Spoofing Attacks on Voice Biometric Systems

Over the last few years, the accuracy of voice recognition systems has increased significantly. However, most organizations utilize these systems either in a limited mode or not at all. That is due to concerns related to information security breaches, with intruders bypassing the system using techniques for imitating a legal user or a bank customer. These fears are fully justified since there are several methods that allow attackers to gain unauthorized access through ASV. The main types of attacks on voice biometric systems are provided below.

Impersonation attacks consist of voice imitation or change by a malefactor without the use of any special devices. It is necessary for the attacker's voice to be similar to the target user's voice to successfully implement the attack. Therefore these security bypass methods have low efficiency. Achieving a match that is close to 100 % is extremely difficult [11] due to individual biometric voice characteristics.

Replay attacks are used to get unauthorized access by falsifying the user's voice with the help of special devices [12]. This attack is implemented by recording the legal user's voice on a device and then replaying it for identification. Its success depends on certain system design features. For example, this type of attack is not effective in text-dependent verification systems and a dynamic passphrase.

The largest threat for ASV is attack methods based on converting the attacker's voice to the user's voice through the use of voice conversion systems. The algorithm includes the stages of training and conversion, where the trained system changes the attacker's voice according to the required parameters. The involvement of the attacker in the process reduces the threat to the ASV.

Speech synthesis attacks are based on text-to-speech technologies. Specially trained systems ensure a natural voice (timbre, smoothness of sound, intonation), the correct placement of signs, stress, and the decoding of special characters. At the verification stage, a synthesized passphrase (for text-dependent systems) or arbitrary speech is created in real-time using the received synthesized voice and an intercepted password. Then this is used to attempt authentication.

Figure 2 shows the different methods of the main types of spoofing attacks on ASV. For example, the most popular method of spoofing attacks is Unit Selection [13–15]. It is implemented in 2 stages. First, a sequence of sound fragments is built considering compliance with the required characteristics. Then it is concatenated to make speech natural. About 8 minutes of a user's speech recording is enough for training modern synthesis systems [16].

The construction of modern ASV should be based on providing a high level of resistance to various types of attacks. Currently, some of these attack methods pose a great threat to such systems. Also, new technologies are being created, and existing ones are improved to perform spoofing attacks. As a result, it is necessary to find solutions and create mechanisms to protect ASV from known and potential threats and increase its reliability for use in the banking sphere.

Methods for Countering Attacks on Voice Biometric Systems

The specifics of the ASV scope in the banking sphere make it necessary to comply with certain requirements. Those include requirements for the entire system (reliability, high recognition accuracy, performance) as well as separate stages, such as pre-processing of biometric features (noise resistance, independence from the speaker). This makes it possible to exclude some combinations of extracted speech features with verification methods. However, it is necessary to study more thoroughly the rest to achieve the optimal system parameters. The most important of them are FAR, FRR, and EER (Equal Error Rate) — the coefficient at which FAR and FRR are the same [17].

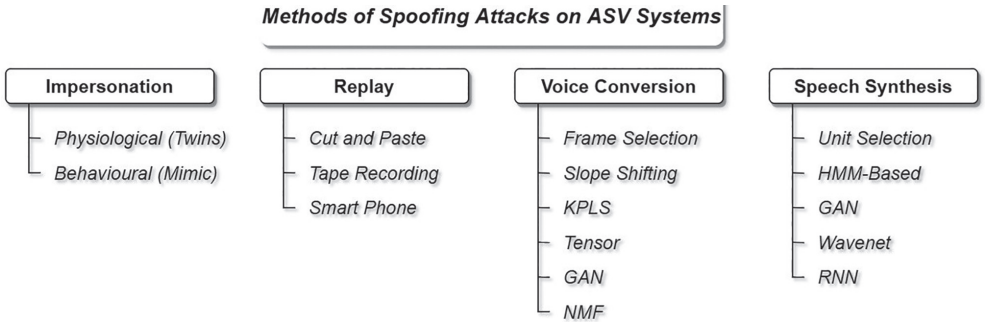

Fig. 2. Existing spoofing attack methods on ASV systems

One solution to increase the security of the ASV involves an attack countering subsystem. In recent years, research has been actively carried out in the field of ASV protection, mostly aimed at identifying the necessary countermeasures and ways to detect spoofing attacks.

There are two main methods of countering replay attacks. The first one is measuring the difference between the channels of recorded and natural speech. Finding recording parameters is also useful to secure ASV.

Attacks based on speech conversion can be detected by natural dynamic variability analysis, which is the characteristic of genuine speech, and voice quality analysis. The latter approach is less dependent on explicit knowledge of the attack but is less efficient.

Countermeasures against speech synthesis attacks are based on searching for phase and tone frequency anomalies. These are known as specific acoustic differences of speech synthesizers, e.g. dynamic ranges of spectral parameters that appear during speech synthesis. Currently, there is no universal solution that is independent of the speech synthesis methods used by attackers.

The Gaussian mixture model, support vector machine, deep neural networks (DNN), and i-vector are common methods of countering the above-mentioned attacks. Different variations of cepstral coefficients are often used for feature extraction. Amplitude, phase, and frequency analysis of the input signal are conducted to identify the inconsistency of natural speech.

There are several methods of acoustic feature extraction. Cepstral voice characteristics are widely used: Mel Frequency Cepstral Coefficent (MFCC), Linear Frequency Cepstral Coefficients (LFCC), Inverse Mel-Frequency Cepstral Coefficients (IMFCC), constant Q phase-based Cepstral Coefficent (CQCC), normalized cosine Cepstral Coefficent (CNPCC), and Cochlear Filter Cepstral Coefficients plus Instantaneous Frequency (CFCCIF).

However, there are other types of feature extraction methods. We could highlight the one based on group delay: Group Delay (GD) or Modified Group Delay (MGD), Modified Group Delay Function Phase Spectrum (MGDFPS), All-Pole Group Delay Function (APGDF). Other feature extraction methods type is based on spectral or frequency characteristics: Log Magnitude Spectrum (LMS), Fundamental Frequency Variation (FFV), Instantaneous Frequency Derivative (IF), Residual Log Magnitude Spectrum (RLMS). Besides these types, it is also worth noting two more methods: Baseband Phase Difference (BPD) and Pitch Synchronous Phase (PSP).

Detailed information about the existing methods of countering attacks and the achieved results are presented in Table 1.

The conducted analysis made it possible to conclude that replay attacks are highly dangerous. The high EER of replay attacks is caused by imperfections of the proposed safety measures for ASV. However, the success of these attacks depends on the quality and quantity of the voice material and how dependent the system is on the text. Replay attacks can be detected by finding identical signal parts for the same phrases. The attacker can attempt to modify certain voice parameters or add noise to the recording. This will negatively impact identification or have no effect at all because of signal pre-processing performed by ASV. Since replay attacks are almost entirely ineffective against text-independent ASV systems, the usage of dynamic passwords is another effective way of countering such attacks. This protection method is already being studied. According to these studies, the GMM method is better suited to short sentences for user verification, but with the increase of sentence length (up to 7–9 seconds), the i-vector method produces lower ERR values [25].

This allows us to conclude that replay technologies and attacks do not have the same potential as voice synthesis. At the current time it is less effective, but for which all of the above restrictions do not apply.

Synthesis technologies have great development potential. They are more difficult to identify and, as a result, pose the greatest danger. Current ASV technologies can barely distinguish natural speech from the synthesized one using hybrid speech synthesis systems based on Unit Selection and HMM technologies [16] and a large training sample (about 4 hours). Usage of such volumes of material remains situational for replay attacks.

Voice synthesis attacks remain a major threat to ASV in the future. Proactive action in this regard is one of the most important objectives of protection systems.

Basics for Building ASV Resistant to Speech Synthesis Attacks

Currently, there is a large number of software solutions for speech synthesis, including free ones[1]. They can build and edit the necessary dictionary databases, support

[1] iSpeech website, Text to Speech for Free — Natural Sounding TTS, https://www.ispeech.org/text.to.speech (date: 25.09.2020).

Table 1. The comparison of existing methods of countering attacks on ASV

Work	Feature extraction method	Classifier	Database*	ERR, %
Replay Attack				
[18]	MFCC/LFCC/IMFCC	GMM	**RSR2015 corpus** DS: 120 speakers (60M, 60F) ES: 60 speakers (30M, 30M)	23.55/12.37/7.50
		i-vector/SVM		26.32/17.37/10.92
		DNN		16.45/8.16/8.29
[19]	MFCC	SVM	**Own dataset** CS1: 20 speakers CS2: 10 speakers More details in [18]	5/3.75/3.12
Synthesis speech attack				
[20]	CQCC/APGDF/FFV	GMM	**ASVspoof 2015 dataset** TS: 25 speakers (10M, 15F) G = 3750, S = 12625 DS: 35 speakers (15M, 20F) G = 3497, S = 49875 ES: 36 speakers (20M, 26F) G = 9404, S = 193404	0.03/0.04/2.02
[21]	MFCC/CNPCC/MFCC-CNPCC	GMM		0.041/5.347/2.694
[22]	LMS/RLMS/GD/MGD/IF/BPD/PSP	Multilayer Perceptron (MLP)		0.543/0.486/0.114/ 1.572/0.428/3.431/1.345
Voice conversion attack				
[23]	MFCC+CFCCIF	GMM	ASVspoof 2015 dataset	1.211
[24]	MGDFPS	SVM/i-vector	NIST SRE2006	8.9

* TS – training set, DS – development set, ES – evaluation set, CS – common set, M – male, F – female, G – genuine, S – spoofed.

technologies to improve synthesized speech, are easy to use, and allow creating customized systems based on the software.

Due to the rapid development of these technologies, identifying and exploiting the weak points of existing voice synthesis technologies is the key to making ASV resistant to speech synthesis. Table 2 shows the main disadvantages of speech synthesis systems that can be used in ASV.

Unpredictable responses of the voice synthesis system to random distortion, adding more information, or anything that the attacker cannot predict is an obvious vulnerability of voice synthesis attacks. However, the construction of an attack detection system based on voice synthesis for ASV cannot be dependent only on the analysis of a large number of accidents, since such protection will not be stable enough.

To build a reliable remote ASV that can be applied in the banking sphere, it is necessary to develop a stable system. Such a system should be based on the following possible solutions, the complex application of which together gives the required result.

Table 2. The main disadvantages of speech synthesis systems that can be used in ASV

Disadvantage	Cause	Synthesized speech feature
Processing information entered into the synthesis system	The complexity of the initial data processing by the synthesis system (normalization, removal of homonymy/homography, recognition of original speech in “voice conversion” systems, etc.) and low dependency on language, such as Russian.	Violation of data flow integrity and decrease in the quality of synthesized speech
Delay in processes	The system needs to understand the context of the pronounced information for high-quality reproduction of the synthesized voice. Moreover, the transformation processes themselves take time (especially in “voice conversion” systems). The system does not start speech synthesis until data input is complete or as directed by the attacker.	Presence of unnatural delays
Interference, noise, distortion while entering information	Random errors in typing or the synthesis system itself, sticky keyboard keys, random sound in the background (during morphing) can cause the synthesis system to function incorrectly and lead to a decrease in the quality of synthesized speech. Increasing the sampling rate of voice samples used for speech synthesis may lead to anomalies such as speckle-noise at the edge of merging synthesized sound elements (phonemes, syllables, words, etc.).	Distortions of the input signal, anomalies (including speckle noise) at the edge of merging synthesized sound elements
Abrupt interruptions in the stream of synthesized voice	Random, unexpected questions of the operator, external factors, for example, another voice in the background registered in the morphing, insertion of extraneous sounds.	Abrupt interruptions and/or swings in the synthesized voice stream
Lack of intonation and emotional dimension	The emotional component of various texts is unique, and the machine is unable to unequivocally interpret them and recognize emotions.	The lack of emotional dimension or intonation

First of all, it is necessary to consider the emotional states of bank customers. Currently, there is no voice synthesis technology capable of imitating a person in a certain psychophysiological state. Emotional state consideration requires the use of voice cepstral characteristics. Secondly, the stability and increased security of ASV rely on regularly adjusting the user's voice fingerprint to maintain its relevance. Additionally, signal and speech intelligibility analysis is required. The signal should not be too smooth, contain unnatural noises, sharp interruptions, changes in the signal level, including point changes, for example, in a phoneme. Reverse Speech Recognition must accurately identify the information reported by the customer.

Generating a dynamic database of passwords containing symbols that are difficult to synthesize and pronounce (abbreviations, homonyms, numbers, and so on) is required to protect ASV against replay attacks. Lastly, it is necessary to understand the language's semantic features (the correct semantic load of speech structures, conjugations, inflections) and conduct appropriate speech analysis to determine the content of possible anomalies.

Conclusion

The paper considers an important problem for information security in the banking sphere that involves countering attacks on voice biometric systems. This problem is quite new, urgent, and difficult to solve.

A description of the main types of attacks on voice biometric systems is presented. These attacks include those based on impersonation, reproduction, speech conversion, speech synthesis. Arguments pointing toward the danger of attacks based on speech synthesis are given.

The existing methods of countering attacks on voice biometric systems are considered. It is concluded that counter-measures to voice synthesis attacks can be based on searching for phase and tone frequency anomalies that appear during speech synthesis. Also it can be based on a preliminary knowledge of the acoustic differences of specific speech synthesizers (for example, the dynamic ranges of spectral parameters). Methods of countering attacks can be based on Gaussian mixture models, support vector machines, deep neural networks, and i-vectors. Various variations of cepstral coefficients are often used as an extractable feature. Amplitude, phase, and frequency analysis of the input signal are carried out to identify inconsistencies compared to natural speech. It is noted that currently there is no universal solution that does not depend on the speech synthesis methods used by the attacker.

The analysis of speech synthesis technologies made it possible to determine the most promising methods of detecting attacks. The following methods can be used as a basis for building voice biometric systems: considering the emotional state of a bank customer (currently, there is no voice synthesis technology capable of imitating a person in a certain psychophysiological state); cepstral voice characteristics; regular adjustment of the user's voiceprint to maintain its relevance; signal analysis (it should not be too smooth, contain unnatural noises, sharp interruptions, changes in the signal level); intelligibility of speech; formation of a dynamic passwords database containing words that are difficult to synthesize and pronounce; identification of semantic speech features (correct semantic load of speech, correct use of cases, conjugations, endings).

Further studies can be done through research on the issues of identifying and countering speech synthesis attacks on voice biometric systems, as well as the development of counter-measures and their software implementation.

References

1. *Fingerprint falsification — possible, but difficult*. Available at: https://www.kaspersky.ru/blog/sas2020-fingerprint-cloning/28101 (accessed: 20.12.2020)
2. Schemelinin V.L. *Methods and complex of means to assess the efficiency of authentication by voice biometric systems*. Dissertation for the degree of candidate of technical sciences, St. Petersburg, 2015. (in Russian)
3. Garipov I.M., Sulavko A.E., Kuprik I.A. Personality recognition methods based on analysis of the characteristics of the outer ear (review). *Information security questions*, 2020, no. 1(128), pp. 33–41. (in Russian)
4. Sudjenkova A.V. Overview of methods for extracting acoustic speech features in speaker recognition. *Transaction of scientific papers of the Novosibirsk state technical university*, 2019, no. 3-4(96), pp. 139–164. (in Russian). doi: 10.17212/2307-6879-2019-3-4-139-164
5. Paul D., Pal M., Saha G. Spectral features for synthetic speech detection. *IEEE Journal of Selected Topics in Signal Processing*, 2017, vol. 11, no. 4, pp. 605–617. doi: 10.1109/JSTSP.2017.2684705
6. Huang T., Wang H., Chen Y., He P. GRU-SVM model for synthetic speech detection. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020, vol. 12022, pp. 115–125. doi: 10.1007/978-3-030-43575-2_9
7. Yang J., Das R.K., Li H. Significance of subband features for synthetic speech detection. *IEEE Transactions on Information Forensics and Security*, 2020, vol. 15, pp. 2160–2170. doi: 10.1109/TIFS.2019.2956589

Литература

1. Подделка отпечатков пальцев — можно, но сложно [Электронный ресурс]. URL: https://www.kaspersky.ru/blog/sas2020-fingerprint-cloning/28101 (дата обращения: 20.12.2020)
2. Щемелинин В.Л. Методика и комплекс средств оценки эффективности аутентификации голосовыми биометрическими системами: диссертация на соискание ученой степени кандидата технических наук / НИУ ИТМО. СПб., 2015.
3. Гарипов И.М., Сулавко А.Е., Куприк И.А. Методы распознавания личности на основе анализа характеристик наружного уха (обзор) // Вопросы защиты информации. 2020. № 1(128). С. 33–41.
4. Судьенкова А.В. Обзор методов извлечения акустических признаков речи в задаче распознавания диктора // Сборник научных трудов Новосибирского государственного технического университета. 2019. № 3-4(96). С. 139–164. doi: 10.17212/2307-6879-2019-3-4-139-164
5. Paul D., Pal M., Saha G. Spectral features for synthetic speech detection // IEEE Journal of Selected Topics in Signal Processing. 2017. V. 11. N 4. P. 605–617. doi: 10.1109/JSTSP.2017.2684705
6. Huang T., Wang H., Chen Y., He P. GRU-SVM model for synthetic speech detection // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2020. V. 12022. P. 115–125. doi: 10.1007/978-3-030-43575-2_9
7. Yang J., Das R.K., Li H. Significance of subband features for synthetic speech detection // IEEE Transactions on Information Forensics and Security. 2020. V. 15. P. 2160–2170. doi: 10.1109/TIFS.2019.2956589

8. Sawada K. *A statistical approach to speech synthesis and image recognition based on Hidden Markov Models*. Doctoral dissertation, Nagoya Institute of Technology, 2018.
9. Saratxaga I., Sanchez J., Wu Z., Hernaez I., Navas E. Synthetic speech detection using phase information. *Speech Communication*, 2016, vol. 81, pp. 30–41. doi: 10.1016/j.specom.2016.04.001
10. van Niekerk B., Nortje L., Kamper H. Vector-quantized neural networks for acoustic unit discovery in the ZeroSpeech 2020 challenge. *Proc. 21st Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020, pp. 4836–4840. doi: 10.21437/Interspeech.2020-1693
11. Wu Z., Yamagishi J., Kinnunen T., Hanilçi C., Sahidullah M., Sizov A., Evans N., Todisco M., Delgado H. ASVspoof: the automatic speaker verification spoofing and countermeasures challenge. *IEEE Journal of Selected Topics in Signal Processing*, 2017, vol. 11, no. 4, pp. 588–604. doi: 10.1109/JSTSP.2017.2671435
12. Lavrentyeva G.M., Novoselov S.A., Kozlov A.V., Kudashev O.Yu., Shchemelinin V.L., Matveev Yu.N., De Marsico M. Audio-replay attacks spoofing detection for speaker recognition systems. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2018, vol. 18, no. 4, pp. 428–437. (in Russian). doi: 10.17586/2226-1494-2018-18-3-428-436
13. Hunt A.J., Black A.W. Unit selection in a concatenative speech synthesis system using a large speech database. *Proc. of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 1996, pp. 373–376. doi: 10.1109/ICASSP.1996.541110
14. Jiang Y., Zhou X.C., Hu Ding Y.J., Ling Z.H., Dai L.R. The USTC system for Blizzard Challenge 2018. *Blizzard Challenge Workshop*, 2018.
15. Kaliev A., Rybin S.V. Speech synthesis: past and present. *Computer Tools in Education*, 2019, no. 1, pp. 5–28. (in Russian). doi: 10.32603/2071-2340-2019-1-5-28
16. Schemelinin V.L., Simonchik K.K. Study of voice verification system tolerance to spoofing attacks using a text-to-speech system. *Journal of Instrument Engineering*, 2014, vol. 57, no. 2, pp. 84–88. (in Russian)
17. Sushchenok O.A. Estimation of an overall performance of biometric systems. *Information Processing Systems*, 2011, no. 4, pp. 79–81. (in Russian)
18. Wu Z., Gao S., Cling E.S., Li H. A study on replay attack and anti-spoofing for text-dependent speaker verification. *Proc. of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2014*, 2014, pp. 7041636. doi: 10.1109/APSIPA.2014.7041636
19. Villalba J., Lleida E. Preventing replay attacks on speaker verification systems. *Proc. of the IEEE International Carnahan Conference on Security Technology, ICCST*, 2011, pp. 06095943. doi: 10.1109/CCST.2011.6095943
20. Pal M., Paul D., Saha G. Synthetic speech detection using fundamental frequency variation and spectral features. *Computer Speech & Language*, 2018, vol. 48, pp. 31–50. doi: 10.1016/j.csl.2017.10.001
21. Alam M.J., Kenny P., Bhattacharya G., Stafylakis T. Development of CRIM system for the automatic speaker verification spoofing and countermeasures challenge 2015. *Proc. 16th Annual Conference of the International Speech Communication Association, INTERSPEECH'15*, 2015, pp. 2072–2076.
22. Xiao X., Tian X., Du S., Xu H., Chng E.S., Li H. Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge. *Proc. 16th Annual Conference of the International Speech Communication Association, INTERSPEECH'15*, 2015, pp. 2052–2056.
23. Patel T.B., Patil H.A. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. *Proc. 16th Annual Conference of the International Speech Communication Association, INTERSPEECH'15*, 2015, pp. 2062–2066.
24. Correia M.J., Abad A., Trancoso I. Preventing converted speech spoofing attacks in speaker verification. *Proc. 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO*, 2014, pp. 1320–1325. doi: 10.1109/MIPRO.2014.6859772
25. Nayana P.K., Mathew D., Thomas A. Performance comparison of speaker recognition systems using GMM and i-vector methods with PNCC and RASTA PLP features. *Proc. of the International Conference on Intelligent Computing, Instrumentation and Control Technologies, ICICICT 2017*, 2017, pp. 438–443. doi: 10.1109/ICICICT1.2017.8342603

8. Sawada K. A statistical approach to speech synthesis and image recognition based on Hidden Markov Models: doctoral dissertation. Nagoya Institute of Technology, 2018.
9. Saratxaga I., Sanchez J., Wu Z., Hernaez I., Navas E. Synthetic speech detection using phase information // Speech Communication. 2016. V. 81. P. 30–41. doi: 10.1016/j.specom.2016.04.001
10. van Niekerk B., Nortje L., Kamper H. Vector-quantized neural networks for acoustic unit discovery in the ZeroSpeech 2020 challenge // Proc. 21st Annual Conference of the International Speech Communication Association, INTERSPEECH. 2020. P. 4836-4840. doi: 10.21437/Interspeech.2020-1693
11. Wu Z., Yamagishi J., Kinnunen T., Hanilçi C., Sahidullah M., Sizov A., Evans N., Todisco M., Delgado H. ASVspoof: the automatic speaker verification spoofing and countermeasures challenge // IEEE Journal of Selected Topics in Signal Processing. 2017. V. 11. N 4. P. 588–604. doi: 10.1109/JSTSP.2017.2671435
12. Лаврентьева Г.М., Новосёлов С.А., Козлов А.В., Кудашев О.Ю. Щемелинин В.Л., Матвеев Ю.Н., Де Марсико М. Методы детектирования спуфинг-атак повторного воспроизведения на голосовые биометрические системы // Научно-технический вестник информационных технологий, механики и оптики. 2018. Т. 18. № 3. С. 428–437. doi: 10.17586/2226-1494-2018-18-3-428-436
13. Hunt A.J., Black A.W. Unit selection in a concatenative speech synthesis system using a large speech database // Proc. of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP. 1996. P. 373–376. doi: 10.1109/ICASSP.1996.541110
14. Jiang Y., Zhou X.C., Hu Ding Y.J., Ling Z.H., Dai L.R. The USTC system for Blizzard Challenge 2018 // Blizzard Challenge Workshop. 2018.
15. Калиев С., Рыбин С.В. Синтез речи: прошлое и настоящее // Компьютерные инструменты в образовании. 2019. № 1. С. 5–28. doi: 10.32603/2071-2340-2019-1-5-28
16. Щемилинин В.Л., Симончик К.К. Исследование устойчивости голосовой верификации к атакам, использующим систему синтеза // Известия высших учебных заведений. Приборостроение. 2014. Т. 57. № 2. С. 84–88.
17. Сущенок О.А. Оценка эффективности работы биометрических систем // Системи обробки інформації. 2011. № 4. С. 79–81.
18. Wu Z., Gao S., Cling E.S., Li H. A study on replay attack and anti-spoofing for text-dependent speaker verification // Proc. of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2014. 2014. P. 7041636. doi: 10.1109/APSIPA.2014.7041636
19. Villalba J., Lleida E. Preventing replay attacks on speaker verification systems // Proc. of the IEEE International Carnahan Conference on Security Technology, ICCST. 2011. P. 06095943. doi: 10.1109/CCST.2011.6095943
20. Pal M., Paul D., Saha G. Synthetic speech detection using fundamental frequency variation and spectral features // Computer Speech & Language. 2018. V. 48. P. 31–50. doi: 10.1016/j.csl.2017.10.001
21. Alam M.J., Kenny P., Bhattacharya G., Stafylakis T. Development of CRIM system for the automatic speaker verification spoofing and countermeasures challenge 2015 // Proc. 16th Annual Conference of the International Speech Communication Association, INTERSPEECH'15. 2015. P. 2072–2076.
22. Xiao X., Tian X., Du S., Xu H., Chng E.S., Li H. Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge // Proc. 16th Annual Conference of the International Speech Communication Association, INTERSPEECH'15. 2015. P. 2052–2056.
23. Patel T.B., Patil H.A. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech // Proc. 16th Annual Conference of the International Speech Communication Association, INTERSPEECH'15. 2015. P. 2062–2066.
24. Correia M.J., Abad A., Trancoso I. Preventing converted speech spoofing attacks in speaker verification // Proc. 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO. 2014. P. 1320–1325. doi: 10.1109/MIPRO.2014.6859772
25. Nayana P.K., Mathew D., Thomas A. Performance comparison of speaker recognition systems using GMM and i-vector methods with PNCC and RASTA PLP features // Proc. of the International Conference on Intelligent Computing, Instrumentation and Control Technologies, ICICICT 2017. 2017. P. 438–443. doi: 10.1109/ICICICT1.2017.8342603

Authors

Aleksandr Yu. Kuznetsov — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57195326635, al.ur.kouznetsov@gmail.com, https://orcid.org/0000-0002-5702-3786

Roman A. Murtazin — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, murtazinroman3161@gmail.com, https://orcid.org/0000-0003-3669-7586
Ilnur M. Garipov — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, i_garipov@mail.ru, https://orcid.org/0000-0003-3108-5484
Evgeny A. Fedorov — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation; Technical Specialist, Laboratory PPS Ltd, Saint Petersburg, 199178, Russian Federation, fyodorov.1997@gmail.com, https://orcid.org/0000-0003-2911-5509

Anna V. Kholodenina — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, annaholodenina@gmail.com, https://orcid.org/0000-0003-1911-3710
Alisa A. Vorobeva — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57191359167, vorobeva@itmo.ru, https://orcid.org/0000-0001-6691-6167

Авторы

Кузнецов Александр Юрьевич — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57195326635, al.ur.kouznetsov@gmail.com, https://orcid.org/0000-0002-5702-3786
Муртазин Роман Андреевич — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, murtazinroman3161@gmail.com, https://orcid.org/0000-0003-3669-7586
Гарипов Ильнур Мидхатович — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, i_garipov@mail.ru, https://orcid.org/0000-0003-3108-5484
Фёдоров Евгений Андреевич — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; технический специалист, ООО «Лаборатория ППШ», Санкт-Петербург, 199178, Российская Федерация, fyodorov.1997@gmail.com, https://orcid.org/0000-0003-2911-5509
Холоденина Анна Викторовна — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, annaholodenina@gmail.com, https://orcid.org/0000-0003-1911-3710
Воробьева Алиса Андреевна — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57191359167,vorobeva@itmo.ru, https://orcid.org/0000-0001-6691-6167

Received 21.12.2020
Approved after reviewing 19.01.2021
Accepted 02.02.2021

Статья поступила в редакцию 21.12.2020
Одобрена после рецензирования 19.01.2021
Принята к печати 02.02.2021