Antispoofing Countermeasures in Modern Voice Authentication Systems

Michael V. Evsyukov [1], Michael M. Putyato [1] and Alexander S. Makaryan [1]

[1] *Kuban State Technological University, 2, Moskovskaya Str., Krasnodar, 350072, Russia*

Abstract

The article provides an overview of modern voice authentication systems. The relevance and importance of voice authentication systems are substantiated. This article focuses on spoofing methods and antispoofing countermeasures because the main challenge in developing voice authentication systems is protection against spoofing. The most commonly used approaches to spoofing, their efficiency, and most important features are described. The existing types of countermeasures against spoofing are presented. The experience of their use against various types of spoofing is described, and links to specific implementations are provided. Improved classification of antispoofing countermeasures is proposed. Promising areas of research in the field of voice authentication are highlighted. According to the authors, the promising areas of research are countermeasures with high generalizing ability, countermeasures specialized against replay attacks, text-dependent countermeasures, development and improvement of joint approaches to assessing the efficiency of an automatic speaker verification system with integrated countermeasures.

Keywords 1

biometrics, authentication, voice, spoofing, information security, artificial intelligence, GMM, SVM.

1. Introduction

According to Google data, 500 million people use Google Assistant monthly [1]. Apple claims the voice assistant Siri handles 25 billion requests every month. [2]. Ease of use and time savings are the main reasons why voice assistants are gaining popularity. In addition, the emergence of a wide range of smart devices and the rapid development of the Internet of Things (IoT) make the voice interface even more in demand, as it can provide the most comfortable user experience. Voice control is implemented, for example, in the Yandex.Station smart speaker, Tesla cars, as well as in various smart home systems.

Thus, voice assistants have entered the daily life of numerous users, and the next natural step is their introduction to payment systems and banking. The main driver for the development of voice solutions is personalization because voice interaction can provide valuable information about the needs and behavior of customers. It allows banks and FinTech companies to offer services that precisely meet the expectations of a particular user.

In a 2017 survey by Business Insider Intelligence in the United States, 8% of respondents said that they used voice commands to buy goods, pay bills, and perform P2P transactions. According to their forecast, by 2022 the number of users of voice interfaces is expected to reach 31% of the US adult population [3].

The graph of growth in the millions of US users of voice interfaces is presented in Figure 1.

Proceedings of VI International Scientific and Practical Conference Distance Learning Technologies (DLT–2021), September 20-22, 2021, Yalta, Crimea

EMAIL: michael.evsyukov@gmail.com (A. 1); putyato.m@gmail.com (A. 2); msanya@yandex.ru (A. 3)

ORCID: 0000-0001-7101-6251 (A. 1); 0000-0003-0414-6034 (A. 2); 0000-0002-1801-6137 (A. 3)

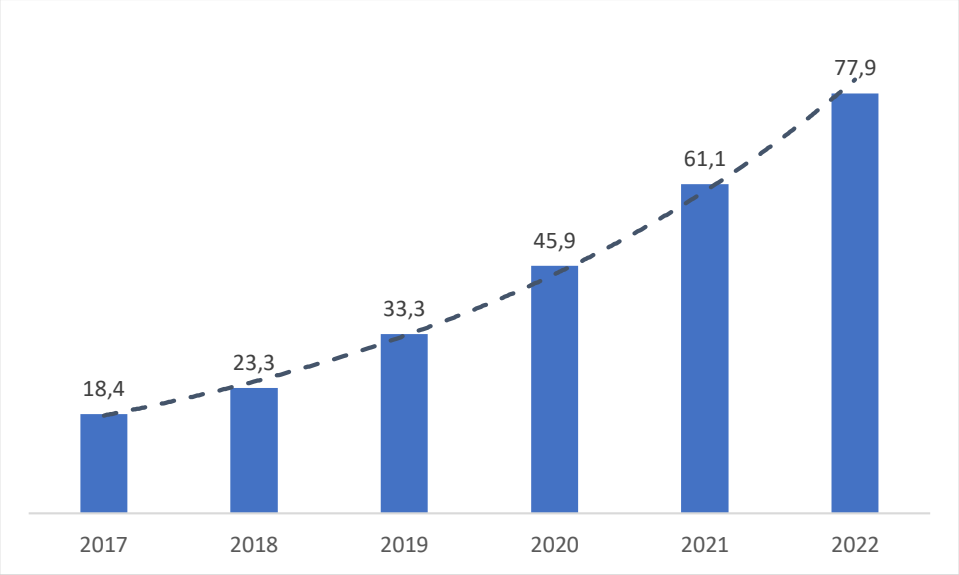

Figure 1: The graph of growth in the millions of US users of voice interfaces

The high consumer value of voice payments is driving banks and payment providers such as PayPal, Amazon, Apple, and Google to develop artificial intelligence technologies dedicated to voice processing.

However, information security problems are the main obstacle that prevents voice payments from gaining the full confidence of banks and users. For them to become as natural as interaction with a merchant or a bank employee, it is necessary to improve the existing methods of protection and authentication [3].

Automatic speaker verification algorithms are well studied, easy to use, and applicable for both continuous and one-time authentication. However, due to the widespread availability of inexpensive audio recording and reproducing devices, they are susceptible to spoofing, i.e. vulnerable to the actions of intruders aimed at impersonating another person. In this regard, the study of antispoofing methods is the main direction in the development of voice authentication systems.

The purpose of this article is to review the current state of research in the field of antispoofing countermeasures for voice authentication systems.

2. Types of Spoofing

Spoofing refers to the actions of an attacker aimed at successfully authenticating in the system under the guise of another person. Due to the widespread availability of high-quality sound recording and reproducing equipment, voice authentication systems are highly susceptible to spoofing.

The main types of spoofing are [4]:

1. Impersonation

This type of spoofing involves one person mimicking the vocal characteristics of another. Impersonation differs from other types of spoofing in that the attacker does not need to use auxiliary technical means and methods to implement it. In this regard, counteraction to this type of spoofing does not require additional countermeasures and is performed by an automatic speaker verification system itself.

2. Speech recording (replay attack)

Speech recording is a simple and efficient form of spoofing. According to a large number of researchers, it poses the most serious threat to automatic speaker verification systems. Its implementation consists in recording a fragment of a person's speech and then replaying it to an authentication system during verification.

3. Voice conversion

Voice conversion involves the use of specialized software that modifies a person's voice in such a way that it becomes similar to another person's voice.

Estimating resistance of various countermeasures against this type of spoofing was the subject of the ASVSpoof 2015 competition [5].

The following speech conversion algorithms were used during ASVSpoof 2015:

- exemplar-based unit selection for voice conversion utilizing temporal information [6];
- adjusting the first Mel-cepstral coefficient to make the voice specter resemble that of the target (one of the simplest algorithms) [7];
- voice conversion based on Gaussian Mixture Model (the most commonly used) [8];
- voice conversion based on the tensor representation of speaker space [9];
- voice conversion, using dynamic kernel partial least squares regression [10].

4. Speech synthesis

This method involves the generation of artificial speech based on an arbitrary text that resembles the voice of a certain person.

Estimating the ability of various countermeasures to resist this type of spoofing was the subject of the ASVSpoof 2015. The following speech synthesis algorithms were used during the competition:

- unit-selection concatenative speech synthesis (this type of spoofing was found to be the most efficient one) [11];
- statistical speech synthesis based on the Hidden Markov Model [12].

Deepfake technology, which implies the use of generative-adversarial neural networks, is a promising method for speech synthesis and voice conversion. The assessment of the ability of countermeasures to resist deepfake-based spoofing will be carried out during the ASVSpoof 2021 competition [13].

5. Disguised attacks on speech processing systems that exploit the human perception of sound

The study [14] considers 4 ways of transforming a voice recording in such a way that it becomes unintelligible to a person, but so that its essential acoustic vocal features remain unchanged, and the recording can pass a speaker verification system or be processed by a speech recognition system.

3. Antispoofing Countermeasures

To resist various spoofing technologies, the voice authentication system must include an additional antispoofing mechanism called countermeasure. The purpose of the countermeasure is to notice the fact that the system is undergoing a spoofing attack.

A general classification of countermeasures against spoofing is presented in [4]. We offer an updated version of it.

1. Challenge-response-based countermeasures

Challenge-response-based countermeasures involve explicit user interaction with the system during authentication. As a rule, when implementing them, the system generates random text that needs to be read by the user. To authenticate the user, a text-independent verification algorithm is used, and the correctness of reading the text is checked by a speech recognition algorithm.

This type of countermeasure is highly effective against the most dangerous type of spoofing – replay attacks. This is because the attacker, as a rule, cannot pre-record the user's speech in such a way that it would be possible to quickly compose a random utterance from its fragments.

2. Acoustic-features-based countermeasures for detection of synthesized and converted speech

This type of countermeasure aims at extracting imperfections from voice recordings that indicate that a piece of speech has been obtained using speech synthesis or voice conversion techniques. Such countermeasures were the subject of research during the ASVSpoof 2015 competition [5].

Countermeasures of this type, like speaker verification methods, are mainly based on the use of short-term spectral characteristics. The most widely used classifiers for these countermeasures are the Gaussian Mixture Model, Support Vector Machines, and Artificial Neural Networks.

3. Voice liveness detection methods based on features of the human vocal tract

Since spoofing involves using loudspeakers, the task of voice authentication can be represented as a combination of the following two tasks:

- authentic a person by his voice characteristics (verification);
- confirm that the source of the voice is a live person (countermeasure).

This type of countermeasure relies on the characteristics of the human vocal tract that cause acoustic effects that are difficult to record and reproduce using artificial means.

The examples of such countermeasures are:

- voice liveness detection algorithms based on pop noise caused by the human breath [15];
- phoneme localization-based liveness detection for voice authentication on smartphones [16].

The scheme of phoneme localization-based liveness detection for voice authentication on smartphones is presented in Figure 2 [16].

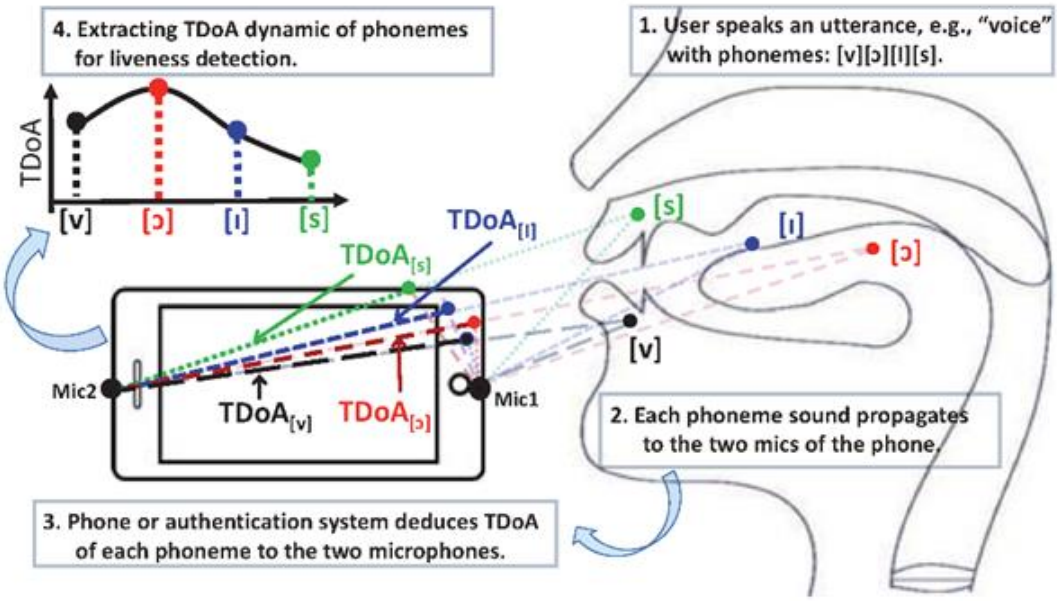

Figure 2: Phoneme localization-based liveness detection for voice authentication on smartphones

- articulatory gesture-based liveness detection [17].

4. Voice liveness detection methods based on features of loudspeakers

At the moment, a wide range of users have access to sound recording and playback devices capable of copying a human voice with very high quality, which continues to evolve. When using the previously described voice characteristics (for example, MFCC), it is extremely difficult to distinguish an artificial voice from a live one, as evidenced by the results of the ASVSpoof 2017 competition [18].

In this regard, it is proposed to use discriminative features of another nature that make it possible to understand if the source of a voice during an authentication attempt is a loudspeaker. For example, in [19] it is proposed to use a magnetic field to distinguish a loudspeaker from a live person.

5. Multi-modal biometric-based methods

This group of methods implies increasing the efficiency of the authentication system and resistance to spoofing by using two or more unrelated biometric characteristics.

For example, work [20] proposes a bimodal identity verification system using Gaussian Mixture Model with Universal Background Model for voice authentication and a face verification system using Gabor's features and linear discriminant analysis. Also, in [21] the possibility of using keyboard handwriting in conjunction with other types of authentication is considered.

4. Conclusion

Currently, there is a large number of ways to perform spoofing against a voice authentication system. On the other hand, there is also a variety of anti-spoofing approaches. Different types of antispoofing are differently efficient against certain types of spoofing. For example, the challenge-response approach works well against voice replay attacks, but it does not apply to other types of spoofing. In turn, countermeasures that rely on the search for acoustic imperfections of the voice used for spoofing are effective against speech synthesis and voice conversion, but much less effective against replay attacks.

In addition, different systems perform differently in countering different spoofing algorithms of the same type.

Therefore, a promising area of research is countermeasures with a high generalizing ability and countermeasures specialized against replay attacks.

In addition, during previous ASVSpoof competitions, only text-independent countermeasures were considered, however, the development of text-dependent countermeasures can have certain advantages as well.

Another promising area of research is the development and improvement of joint approaches to assessing the efficiency of automatic speaker verification systems with integrated countermeasures.

5. References

[1] L. Eadicicco, Google just revealed that half a billion people around the world are using the Google Assistant as it battles with Amazon to conquer the smart home, 2020. URL: https://www.businessinsider.com/google-assistant-500-million-users-challenges-amazon-alexa-2020-1

[2] B. Kinsella, Apple Still in Holding Pattern on Voice, Siri Used 25 Billion Times Per Month But New Features Limited, 2020: URL: https://voicebot.ai/2020/06/22/apple-still-in-holding-pattern-on-voice-siri-used-25-billion-times-per-month-but-new-features-limited/

[3] D.V. Dyke, Soon nearly a third of US consumers will regularly make payments with their voice, 2017. URL: https://www.businessinsider.com/the-voice-payments-report-2017-6?r=US&IR=T

[4] B. Hao, X. Hei, Voice Liveness Detection for Medical Devices, in: D.R. Kisku (Ed.), P. Gupta (Ed.), J.K. Sing (Ed.), Design and Implementation of Healthcare Biometric Systems, 2019, pp. 109-136. DOI: 10.4018/978-1-5225-7525-2.ch005

[5] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilc, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, ASVspoof: the Automatic Speaker Verification Spoofing and Countermeasures Challenge, IEEE Journal of Selected Topics in Signal Processing, 6(1) (2016) 588-604. DOI: 10.1109/JSTSP.2017.2671435

[6] Z. Wu, T. Virtanen, T. Kinnunen, E. Chng, H. Li, Exemplar-based unit selection for voice conversion utilizing temporal information, in: Proceedings of 14th Annual Conference of the International Speech Communication Association, Interspeech 2013, 2013, pp. 3057-3061.

[7] T. Fukada, K. Tokuda, T. Kobayashi, S. Imai, An adaptive algorithm formel-cepstral analysis of speech, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing, ICASSP-92, 1992, pp. 137-140. DOI: 10.1109/ICASSP.1992.225953

[8] T. Toda, A.W. Black, K. Tokuda, Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory, IEEE Transactions on Audio, Speech, and Language Processing 15(8) (2007) 2222-2235. DOI: 10.1109/TASL.2007.907344

[9] D. Saito, K. Yamamoto, N. Minematsu, K. Hirose, One-to-many voice conversion based on the tensor representation of speaker space, in: 12th Annual Conference of the International Speech Communication Association, INTERSPEECH 2011, Florence, Italy, 2011, pp. 653-656.

[10] E. Helander, H. Sil´en, T. Virtanen, M. Gabbouj, Voice conversion using dynamic kernel partial least squares regression, IEEE Transactions on Audio Speech and Language Processing 20(3) (2012) 806-817. DOI: 10.1109/TASL.2011.2165944

[11] A.J. Hunt, A.W. Black, Unit selection in a concatenative speech synthesis system using a large speech database, in: IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, 1996. DOI: 10.1109/ICASSP.1996.541110

[12] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, J. Isogai, Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained smaplr adaptation algorithm, IEEE Trans. Audio, Speech and Language Processing, 17(1) 2009 66-83. DOI: 10.1109/TASL.2008.2006647

[13] H. Delgado, N. Evans, T. Kinnunen, K.A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang, J. Yamagishi, ASVspoof 2021: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan, ASVspoof consortium, 2021. URL: https://www.asvspoof.org/asvspoof2021/asvspoof2021_evaluation_plan.pdf

[14] H. Abdullah, W. Garcia, C. Peeters, P. Traynor, K. Butler, J. Wilson, Practical Hidden Voice Attacks against Speech and Speaker Recognition Systems, The Network and Distributed System Security Symposium (NDSS), 2019. URL: https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019_08-1_Abdullah_paper.pdf
[15] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, T. Matsui, Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification, in 16th Annual Conference of the International Speech Communication Association, Interspeech 2015, 2015. pp. 239-243. DOI: 10.21437/Interspeech.2015-92
[16] L. Zhang, S. Tan, J. Yang, et al., VoiceLive: A phoneme localization-based liveness detection for voice authentication on smartphones, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, CCS'16, 2016, pp. 1080-1091. DOI: 10.1145/2976749.2978296
[17] L. Zhang, S. Tan, J. Yang, Hearing your voice is not enough: An articulatory gesture-based liveness detection for voice authentication, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, CCS'17, 2017, pp. 55-71. DOI: 10.1145/3133956.3133962
[18] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, K.A. Lee, 19th Annual Conference of the International Speech Communication Association, Interspeech 2018, Stockholm, Sweden, 2018. DOI: 10.21437/Interspeech.2017-1111
[19] L. Li, Y. Chen, D. Wang, et al., A study on replay attack and anti-spoofing for automatic speaker verification, in: Proceedings of 18th Annual Conference of the International Speech Communication Association, Interspeech 2017, Stockholm, Sweden, 2017, pp.92-96.
[20] A. Usoltsev, D. Petrovska-Delacrétaz, K. Houssemeddine, Full Video Processing for Mobile Audio-Visual Identity Verification, in: Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2016, 2016, pp. 552-557.
[21] M.M. Putyato, A.S. Makaryan, S.M. Chich, V.K. Markova, System development for identification and confirmation of access legitimacy based on biometric authentication dynamic methods, Caspian Journal: Control and High Technologies, 3(51) (2020) 83-93.