Vulnerability of Voice Verification Systems to Spoofing Attacks by TTS Voices Based on Automatically Labeled Telephone Speech

Vadim Shchemelinin[1,2], Mariia Topchina[2], and Konstantin Simonchik[2]

[1] National Research University of Information Technologies, Mechanics and Optics, St.Petersburg, Russia
www.ifmo.ru
[2] Speech Technology Center Limited, St.Petersburg, Russia
www.speechpro.com
{shchemelinin,topchina,simonchik}@speechpro.com

Abstract. This paper explores the robustness of a text-dependent voice verification system against spoofing attacks that use synthesized speech based on automatically labeled telephone speech. Our experiments show that when manual labeling is not used in creating the synthesized voice, and the voice is based on telephone speech rather than studio recordings, False Acceptance error rate decreases significantly compared to high-quality synthesized speech.

Keywords: spoofing, speech synthesis, unit selection, HMM, speaker recognition.

1 Introduction

Information technology plays an increasingly large role in today's world, and different authentication methods are used for restricting access to informational resources, including voice biometrics. Examples of using speaker recognition systems include internet banking systems, customer identification during a call to a call center, as well as passive identification of a possible criminal using a preset "blacklist" [1]. Due to the importance of the information that needs to be protected, requirements for biometric systems are high, including robustness against potential breakins and other attacks. Robustness of the basic technology of voice biometrics has greatly improved in recent years. For instance, the latest NIST SRE 2012 competition [2] showed that the EER of text-independent speaker recognition systems is down to 1.5-2% in various conditions. However, the vulnerability of these systems to spoofing attacks is still underexplored and needs serious examination.

For this reason, a new direction of spoofing [3,4,5], and anti-spoofing in voice biometric system has recently appeared. Different spoofing methods were examined. For example, [6] describes methods based on "Replay attack", "Cut and paste", "Handkerchief tampering" and "Nasalization tampering". However, spoofing using text-to-speech synthesis based on the target speakers voice remains one of the most successful spoofing methods. [7] examines the method of spoofing which is performed using a

A. Ronzhin et al. (Eds.): SPECOM 2014, LNAI 8773, pp. 475–481, 2014.

hybrid TTS method that combines Unit Selection and HMM. The likelihood of false acceptance when using high-quality speech synthesis and a speech database recorded with studio quality can reach 98%.

This paper explores the robustness of a text-dependent verification system against spoofing based on the method described in [7] using a synthesized voice based on automatically labeled "free" speech recorded in the telephone channel. This attack scenario does not require expert knowledge for preparing a synthesized voice and is more likely to be implemented by criminals.

The aim of our research is to find out how strongly False Acceptance (FA) error rate will decrease if the perpetrator cannot access an expert for speech database labeling, and if the database is recorded in the telephone channel.

2 The Voice Verification System

A typical scenario of the functioning of a text-dependent verification system is shown in figure 1. The user connects to the text-dependent verification system and inputs his or her unique ID. The system sends a passphrase for the user to pronounce. The user pronounces the passphrase, the system compares it to the model recorded during user registration and makes the decision whether the user should be allowed or denied access.

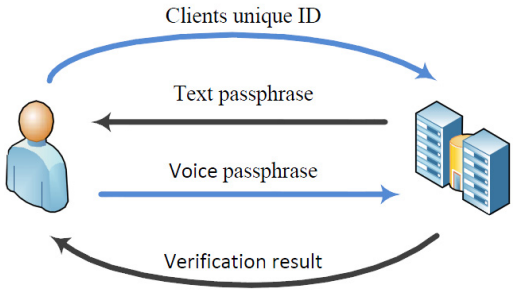

Fig. 1. The process of text-dependent verification

In our experiments we used i-vector based speaker recognition system [8,9].

We used special signal preprocessing module, which included energy based voice activity detection, clipping [10], pulse and multi-tonal detection. The front-end computes 13 mel-frequency cepstral coefficients, as well as the first and second derivatives, to yield a 39 dimensional vector per frame. The derivatives are estimated over a 5-frame context. To obtain these coefficients, speech samples are pre-emphasized, divided into 22ms window frames with a fixed shift of 11ms, and each frame is subsequently multiplied by a Hamming window function.

We also applied a cepstral mean subtraction (CMS) and did not apply Feature Warping [11] for the cepstral coefficients.

We used a gender-independent universal background model (UBM) with 512 - component gaussian mixture model (GMM), obtained by standard ML-training on the telephone part of the NIST's SRE 1998-2010 datasets (all languages, both genders) [12], [13].

In our study we used more than 4000 training speakers in total. We also used a diagonal, not a full-covariance GMM UBM.

The i-vector extractor was trained on more than 60000 telephone and microphone recordings from the NIST 1998-2010 comprising more than 4000 speakers' voices.

The main expression defining the factor analysis of the GMM parameters with the aim of lowering data dimensionality is given below:

$$\mu = m + T\omega + \epsilon,$$

where μ is the supervector of the GMM parameters of the speaker model,
m is the supervector of the UBM parameters,
T is the matrix defining the basis in the reduced feature space,
ω is the i-vector in the reduced feature space, $\omega \in N(0, 1)$,
ϵ is the error vector.
LDA matrix was trained on the same data from the NIST 1998-2010.

3 The Method of Spoofing the Verification System

We chose to model a spoofing attach method based on a TTS (Text-to-Speech) system developed by Speech Technology Center Ltd (STC) [14]. [15] demonstrates that when the synthesized voice is built using 8 minutes of free speech recorded in a studio environment and manually labeled, the spoofer can achieve 44% likelihood of false acceptance (Table 1).

Table 1. FA verification error for spoofing the verification system based on different length of high quality speech with professional labeling (amount of free speech used for passphrase synthesis)

Length of speech data for TTS	FA for threshold in calibration EER point	FA for threshold in calibration $FA = 1\%$ point
1 minute	12.7%	1.5%
3 minutes	34.9%	7.9%
8 minutes	44.4%	19.1%
30 minutes	55.6%	23.8%
4 hours	100%	98.4%

In our experiment we used automatic database labeling, which includes $F0$ period labeling and phone labeling. $F0$ labeling is done by means of the autocorrelation method of $F0$ calculation with preliminary filtering and postprocessing for more precise labeling of $F0$ periods. Low frequency filtering is used to lower the $F0$ detection error by deleting components higher than 500Hz from the signal. High frequency filtering is used to detect the fragments that have no $F0$ (nonvocalized phones).

Phone labeling is done automatically using automatic speech recognition (ASR) modules based on Hidden Markov Models (HMM). The labeling is based on forced alignment of the transcription and the signal. It involves three steps:

1. Building acoustic models of monophones, since monophones are best suited for this task.
2. Obtaining the "ideal" labeling that exactly matches the required transcription, and the "real" labeling that more closely matches the recording.
3. Automatic correction of the obtained phone labels based on the $F0$ labeling.

The process of automatically building a TTS voice is described in detail in [16].

The spoofing attack scheme modeled in this paper is demonstrated in Figure 2. The attack is based on creating a TTS voice based on previously recorded free speech of a verification system user and its automatic segmentation. In the process of text-prompted verification, the text of the passphrase is received and it is then synthesized with the users voice by the spoofing system.

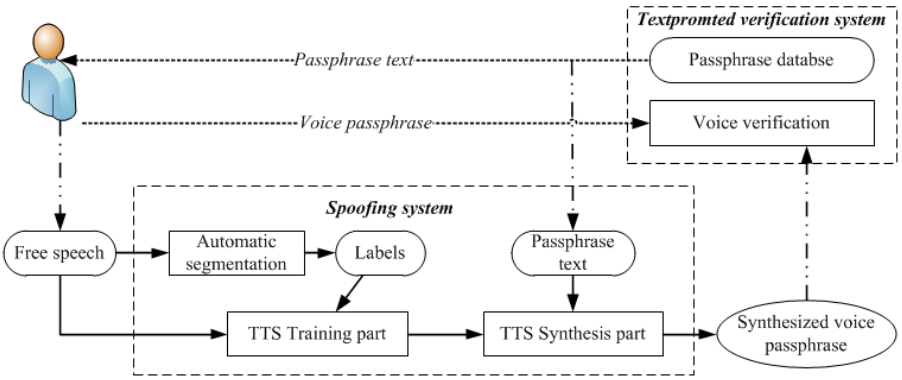

Fig. 2. Scheme of spoofing a text-prompted verification system using TTS technology

As previously recorded free speech we used a Russian phone speech database with 5 speakers whose voices were used for creating a TTS system. Examples of passphrases include: "2014 year", "City of Ekaterinburg, Railway Station street, 22, Railway Station"; "pay three roubles and publish an ad in the bulletin", etc. It is important to note that the recorded phrases were not included in the TTS database. In total, 95 phrases by different speakers were recorded.

The verifications system thresholds were calibrated using a YOHO speech database [17] consisting of 138 speakers (male and female) each of whom pronounced a "Combination lock" phrases of the form "36-24-36", with about 1.5-2 seconds of pure speech. Only one passphrase was used for enrollment and one for the verification. Two verification system thresholds were set:

1. A threshold based on Equal Error Rate (EER), so-called $ThresholdEER$. EER was estimated as 4% on the YOHO database.

2. A threshold with the likelihood of false acceptance not higher than 1% ($Threshold\ FA$). This threshold is usually used in systems where it is necessary to provide maximum defense against criminal access.

Then, for each speaker, attempts to access the system were made using a TTS voice that was created using the speech material of this speaker. The length of speech material used for creating the TTS voice varied from 1 minute to 8 minutes of speech. The experimental results are presented in Figure 3.

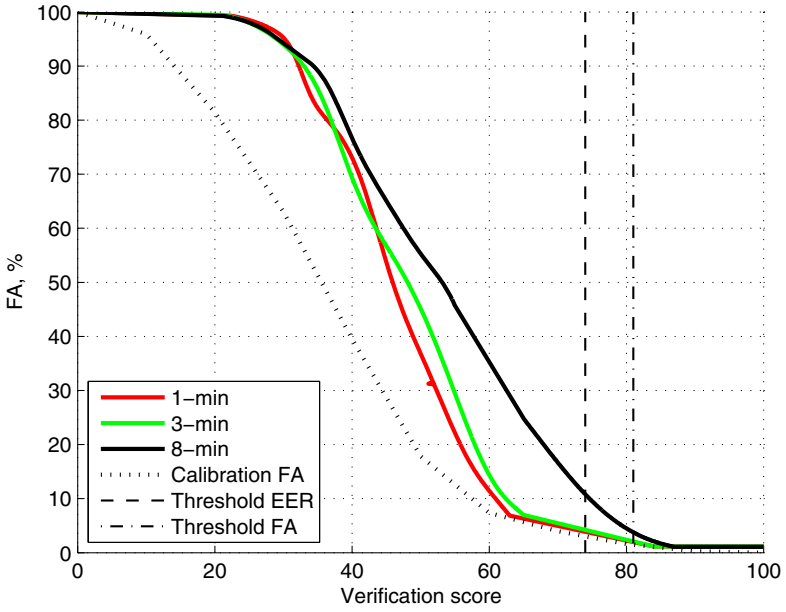

Fig. 3. FA diagrams for spoofing the verification system with a TTS voice based on different durations of telephone speech with automatic labeling (amount of free speech used for passphrase synthesis)

In Table 2, for different verification system thresholds, presented comparisons of the FA values obtained with automatic labeling of free speech with the results showed by [7], where free speech was labeled manually by experts.

As can be seen from the table, if automatically labeled telephone speech is used in TTS, the False Acceptance error rate is strongly decreased.

4 Conclusions

We analyzed the vulnerability of state-of-the-art verification methods against spoofing using a hybrid TTS system based on automatically labeled speech. As demonstrated

Table 2. FA verification error for spoofing the verification system based on different length of high quality speech with professional and automatic labeling (amount of free speech used for passphrase synthesis)

	FA for $ThresholdEER$		FA for $ThresholdFA$	
Length of speech data for TTS	Expert labeling	Automatic labeling	Expert labeling	Automatic labeling
1 minute	12.7%	1.1%	1.5%	1.1%
3 minutes	34.9%	4.6%	7.9%	1.8%
8 minutes	44.4%	10.8%	19.1%	4.5%

by the experiments, spoofing using a TTS voice based on telephone speech that was labeled automatically yields a significantly lower False Acceptance rate compared to a TTS voice based on speech recorded in a studio environment and manually labeled by experts. For instance, when 8 minutes of speech were used for TTS voice creation, the new spoofing method gave only a 10% False Acceptance error, compared to the 44% obtained earlier.

However, our results show once again that it is highly necessary to test verification systems against spoofing by different methods, and to develop anti-spoofing algorithms. Even a 10% False Acceptance error rate, provided the attack is fully automated, makes it possible to easily break into a voice verification system.

Acknowledgments. This work was partially financially supported by the Government of Russian Federation, Grant 074-U01.

References

1. Matveev, Y.: Biometric technologies of person identification by voice and other modalities, Vestnik MGTU. Priborostroenie. Biometric Technologies 3(3), 46–61 (2012)
2. The NIST Year 2012 Speaker Recognition Evaluation Plan, `http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf`
3. Wu, Z., Kinnunen, T., Chng, E.S., Li, H., Ambikairajah, E.: A Study on spoofing attack in state-of-the-art speaker verification: the telephone speech case. In: Proc. of the APSIPA ASC, Hollywood, USA, pp. 1–5 (December 2012)
4. Wu, Z., Kinnunen, T., Chng, E.S., Li, H.: Speaker verification system against two different voice conversion techniques in spoofing attacks, Technical report (2013), `http://www3.ntu.edu.sg/home/wuzz/`
5. Kinnunen, T., Wu, Z., Lee, K.A., Sedlak, F., Chng, E.S., Li, H.: Vulnerability of Speaker Verification Systems Against Voice Conversion Spoofing Attacks: the Case of Telephone Speech. In: Proc. of the ICASSP, Kyoto, Japan, pp. 4401–4404 (March 2012)
6. Villalba, E., Lleida, E.: Speaker verification performance degradation against spoofing and tampering attacks. In: Proc. of the FALA 2010 Workshop, pp. 131–134 (2010)
7. Shchemelinin, V., Simonchik, K.: Examining Vulnerability of Voice Verification Systems to Spoofing Attacks by Means of a TTS System. In: Železný, M., Habernal, I., Ronzhin, A. (eds.) SPECOM 2013. LNCS, vol. 8113, pp. 132–137. Springer, Heidelberg (2013)
8. Kenny, P.: Bayesian speaker verification with heavy tailed priors. In: Proc. of the Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic (June 2010)

9. Simonchik, K., Pekhovsky, T., Shulipa, A., Afanasyev, A.: Supervized Mixture of PLDA Models for Cross-Channel Speaker Verification. In: Proc. of the 13th Annual Conference of the International Speech Communication Association, Interspeech 2012, Portland, Oregon, USA, September 9-13 (2012)
10. Aleinik, S., Matveev, Y., Raev, A.: Method of evaluation of speech signal clipping level. Scientific and Technical Journal of Information Technologies, Mechanics and Optics 79(3), 79–83 (2012)
11. Pelecanos, J., Sridharan, S.: Feature warping for robust speaker verication. In: Proc. of the Speaker Odyssey, the Speaker Recognition Workshop, Crete, Greece (2001)
12. Matveev, Y., Simonchik, K.: The speaker identification system for the NIST SRE 2010. In: Proc. of the 20th International Conference on Computer Graphics and Vision, GraphiCon 2010, St. Petersburg, Russia, September 20-24, pp. 315–319 (2010)
13. Kozlov, A., Kudashev, O., Matveev, Y., Pekhovsky, T., Simonchik, K., Shulipa, A.: Speaker recognition system for the NIST SRE 2012. SPIIRAS Proceedings 25(2), 350–370 (2012)
14. Chistikov, P., Korolkov, E.: Data-driven Speech Parameter Generation for Russian Text-to-Speech System. Computational Linguistics and Intellectual Technologies. In: Annual International Conference "Dialogue", pp. 103–111 (2012)
15. Simonchik, K., Shchemelinn, V.: "STC SPOOFING" Database for Text-Dependent Speaker Recognition Evaluation. In: Proc. of SLTU-2014 Workshop, St. Petersburg, Russia, May 14-16, pp. 221–224 (2014)
16. Solomennik, A., Chistikov, P., Rybin, S., Talanov, A., Tomashenko, N.: Automation of New Voice Creation Procedure For a Russian TTS System. Vestnik MGTU. Priborostroenie, "Biometric Technologies" 2, 29–32 (2013)
17. "YOHO Speaker Verification" database, Joseph Campbell and Alan Higgins, `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC94S16`