"STC SPOOFING" DATABASE FOR TEXT-DEPENDENT SPEAKER RECOGNITION EVALUATION

Konstantin Simonchik[1,2], *Vadim Shchemelinin*[1,2]

[1]The National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems
[2]Speech Technology Center Ltd.

ABSTRACT

The paper describes the "STC Spoofing" database, which consists of a set of recordings of "live" speech by several speakers, as well as synthesized speech fragments obtained using a TTS engine based on these speakers' voices.
The database can be used for testing the robustness of text-dependent speaker verification systems against spoofing attacks, as well as for research and development of methods for fighting break-ins into biometric systems that are performed using synthesized speech.

Index Terms— Database, spoofing, anti-spoofing

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". Due to the importance of the information that needs to be protected, requirements for biometric systems are high, including robustness against potential break-ins 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 [1] 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 [2-4] and anti-spoofing in voice biometric system has recently appeared. Different spoofing methods were examined. For example, [5] 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 speaker's voice remains one of the most successful spoofing methods. [6] examines the method of spoofing which is performed using a hybrid TTS method that combines Unit Selection and HMM. The likelihood of false acceptance when using high-quality speech synthesis can reach 98%.

This paper describes a speech database collected for the purpose of examining the robustness of voice verification systems against spoofing using TTS systems. Section 2 contains a detailed description of the database format, as well as the TTS engine that was used to generate the synthesized part of the database. Section 3 describes testing a text-dependent verification system using the proposed database. In section 4 we draw the conclusions.

2. DATABASE DESCRIPTION

The "STC spoofing" speech database contains Russian speech of 7 speakers (2 men and 5 women). For each speaker, there are several "real" passphrases. For each passphrase, there are its synthesized versions of different quality: from the worst, when only 30 seconds of speech was used for TTS voice building, to the best, when the synthesized voice was based on 3 hours of the speaker's speech.

Examples of passphrases include: "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, 63 phrases by different speakers were recorded.

2.1. The "Human" part

The "human" part of the speech database is a set of "live" model recordings of passphrases by 7 speakers made in the microphone channel. There are 3 phrases for each speaker, so this part comprises 21 recordings.

Full properties of the model recordings are given in Table 1.

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

Table 1. Properties of model recordings.

Parameter	Value
SNR	23-43 dB
File length	2.6 – 7 sec
"Pure" speech length	1.2-3.3 sec
Reverberation	0.15 – 0.4 sec
Sampling frequency	22050 Hz

2.2. The "Synthesis" part

The synthesized part of the speech database consists of recordings obtained using a TTS system developed by Speech Technology Center Ltd (STC) [7]. The synthesized phrases are the same as the "live" model phrases from the first part of the database. This makes it possible to use them for imitating verification attempts by an imposter.

The TTS engine is based on two most popular approaches:

1. The Unit Selection algorithm (speech element selection). This approach makes it possible to synthesize speech with maximum naturalness, given an accurately segmented voice database of a large size (10 hours and more). On the other hand, the second approach, which produces synthesized speech that is less natural, has the advantages presented below.
2. Statistical models (HMM TTS), which produce synthesized speech that is less natural, but smoother, without detectable phone boundaries (pitch or energy leaps), which are usual for concatenative synthesis. In addition, the HMM-based method provides an easy way to modify voice characteristics by using speaker adaptation/interpolation techniques. Finally, applying the HMM-based speech synthesis method makes it possible to create a new TTS voice in much less time and to reduce the memory size required for storing the voice data.

Structurally, the system is divided in two parts (see Figure 1):

- training part (the preparation stage);
- synthesis part.

A training database is created based on the "free speech" containing a set of sound files (each file contains a single recorded sentence) and a set of corresponding label files (these contain information about the speech elements in each sound file) [8-10].

Experiments [7] show that the naturalness of speech synthesized by the hybrid TTS system is increased compared to systems based only on Unit Selection or hidden Markov models. A detailed description of the TTS system is given in [7].

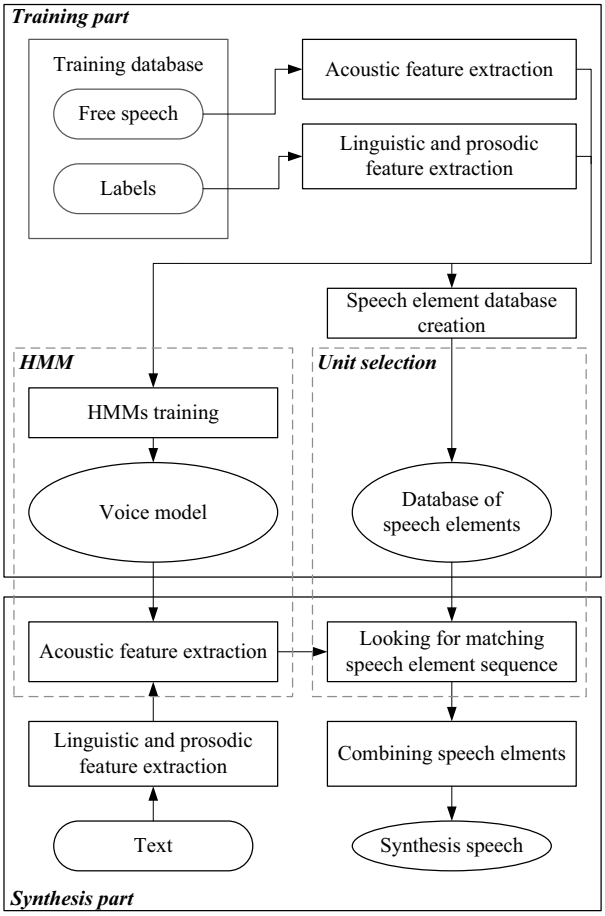

Fig. 1 *Diagram illustrating the basic steps conducted by the speech synthesis engine*

2.3. Speech data format

This section contains a detailed description of the data format of the "STC Spoofing v. 1.0" database.

Database directories:

The *\human* folder contains human speech files with utterances of a fixed passphrase. Total: 7 speakers with 3 phrases for each.

The *\synthesis* folder contains synthesized speech. It is divided into subfolders depending on the duration of speech material used for creating the TTS voice:

- *\30s* - 30 seconds of free speech for TTS creation;
- *\01m* - 1 minute of free speech for TTS creation;
- *\05m* - 5 minutes of free speech for TTS creation;
- *\08m* - 8 minutes of free speech for TTS creation;
- *\10m* - 10 minutes of free speech for TTS creation;
- *\15m* - 15 minutes of free speech for TTS creation;
- *\20m* - 20 minutes of free speech for TTS creation;
- *\30m* - 30 minutes of free speech for TTS creation;
- *\3h* - 3 hours of free speech for TTS creation.

File names in the database have the following format:
GXX_YY_ZZ_TYPE.wav
G - gender: 'M' - male, 'F' - female
XX - speaker ID
YY - passphrase ID
ZZ - session ID
TYPE - type of speech: 'synth' - synthesized speech, 'human' - human speech.

2.4. "STC Spoofing" request

To obtain the database for non-commercial use, please contact the authors of this paper via the email address given in the paper title.

3. EXAMPLE OF USING THE DATABASE

The proposed database was used for testing the robustness of a text-dependent verification system [11]. The spoofing attack scheme modeled in this test is demonstrated in Figure 2. The attack is based on creating a TTS voice based on previously recorded free speech of a verification system customer. In the process of text-prompted verification, the text of the passphrase is received and it is then synthesized with the customer's voice by the spoofing system.

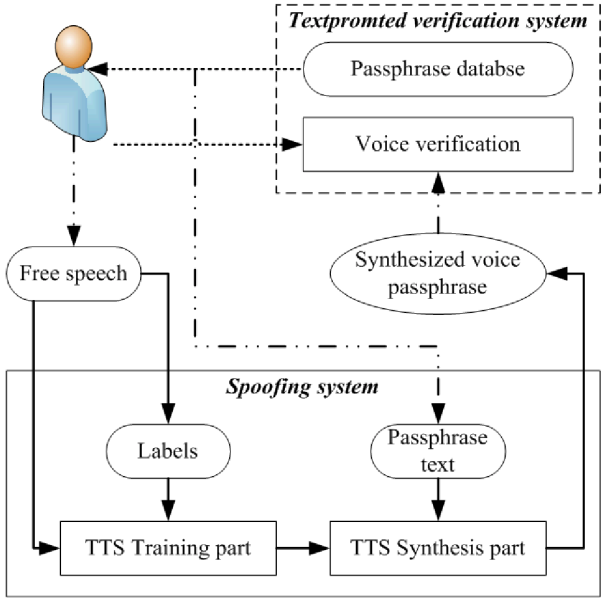

Fig. 2. Scheme of spoofing a text-prompted verification system using TTS technology.

The verifications system thresholds were calibrated using a YOHO speech database [12] 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% (ThresholdFA1). 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 4 hours of speech.

The results of testing robustness against spoofing attacks based on speech synthesis show that the database makes it possible to detect vulnerability of a text-dependent verification system to the proposed attack scheme. False acceptance error graphs for the tested verification system for different folders of the database are given in Figure 3. [11] describes the results in more detail.

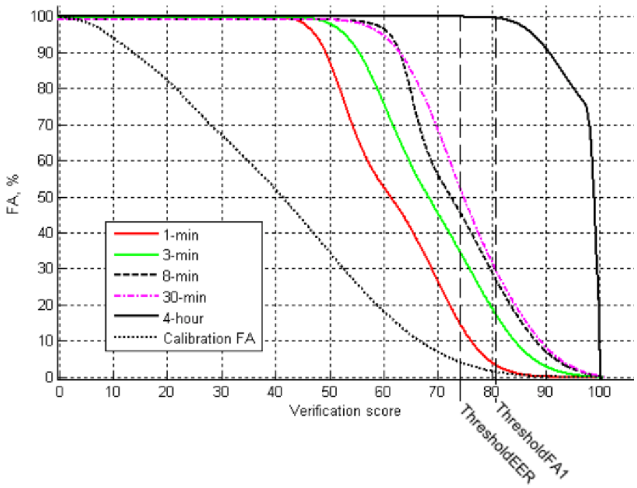

Fig. 3. FA diagrams for spoofing the verification system based on different database directories (volumes of free speech used for passphrase synthesis).

4. CONCLUSIONS

This paper presented a speech database aimed at testing the robustness of verification systems against spoofing attacks, as well as developing methods for fighting biometrical system break-ins. We described the collection of the database and the technology that was used for generating its synthesized part. It should be noted that the current drawback of the database is the small number of speakers and the use of only one TTS technology. We are planning to correct these drawbacks in the near future. We are also open for any suggestions concerning both increasing the amount of speech material and the number of TTS systems used for the generation of the synthesized part of the database.

5. REFERENCES

[1] The NIST Year 2012 Speaker Recognition Evaluation Plan, http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf.

[2] Z. Wu, T. Kinnunen, E.S. Chng, H. Li, E. Ambikairajah, "A Study on spoofing attack in state-of-the-art speaker verification: the telephone speech case", In: Proc. APSIPA ASC 2012, Hollywood, USA, 2012, pp. 1-5.

[3] Z. Wu, E. S. Chng, and H. Li, "Speaker verification system against two different voice conversion techniques in spoofing attacks," Technical report: available at http://www3.ntu.edu.sg/home/wuzz/.

[4] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, H. Li, "Vulnerability of Speaker Verification Systems Against Voice Conversion Spoofing Attacks: the Case of Telephone Speech", In:Proc. ICASSP 2012, Kyoto, Japan, 2012, pp. 4401-4404.

[5] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in FALA 10 workshop, 2010, pp. 131–134.

[6] Shchemelinin V., Simonchik K. Examining Vulnerability of Voice Verification Systems to Spoofing Attacks by Means of a TTS System // In Proc. Speech and Computer SPECOM-2013, Pilsen, Czech Republic, 2013, pp. 132-137.

[7] Pavel Chistikov, Evgeny Korolkov. Data-driven Speech Parameter Generation For Russian Text-to-Speech System. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2012). Issue 11 (18) Volume 1, p. 103-111.

[8] Prodan A., Chistikov P., Talanov A. (2010), Voice building system for Russian TTS system "Vital Voice", Proceedings of the Dialogue-2010 International Conference, No 9 (16), pp. 394-399.

[9] Smirnova N., Chistikov P. (2011), Software for Automated Statistical Analysis of Phonetic Units Frequency in Russian Texts and its Application for Speech Technology Tasks, Proceedings of the Dialogue-2011 International Conference, No 10 (17), pp. 632-643.

[10] Chistikov P., Khomitsevich O. (2011), On-line automatic sentence boundary detection in a Russian ASR system, SPECOM 2011 International Conference, pp. 112-117.

[11] Shchemelinin V., Simonchik K. Study of Voice Verification System Tolerance To Spoofing Attacks Using A Text-To-Speech System //Scientific and Technical Journal «Priborostroenie», 2014, Issue 2, pp. 84-88.

[12] "YOHO Speaker Verification" database, Joseph Campbell and Alan Higgins, http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC94S16.