24th INTERNATIONAL CONGRESS ON SOUND AND VIBRATION
23–27 July 2017, London

WEAKNESSES OF VOICE BIOMETRICS – SPEAKER VERIFICATION SPOOFING USING SPEECH SYNTHESIS

Milan Rusko, Marian Trnka, Sakhia Darjaa and Marian Ritomský

Institute of Informatics of the Slovak Academy of Sciences, Bratislava, Slovakia
email: milan.rusko@savba.sk

It is well known that voice biometric systems are vulnerable to imposture. Typical spoofing attacks that are performed on speaker verification systems can use impersonation, replay, voice conversion, artificial signals, speech synthesis and other approaches. In this work the authors analyze the effectiveness of potential spoofing attacks using different speech synthesis approaches. The state of the art speech synthesizers using deep neural networks for acoustic characteristics modelling and high-quality vocoder with enhanced excitation function model, are able to generate a very natural speech signal keeping the personal characteristics of the target speaker, whose speech samples were used either for adaptation of some general voice, or for the complete design of the speech synthesizer. The synthesis artefacts are nearly imperceptible in their signal. However some systems using unit selection from a big speech database often offer comparable or even better speech quality. An i-vector based speaker verification system with PLDA scoring was used for experiments. Several male and female synthesized voices, both in vocoder-based and unit-selection versions were tested for their spoofing effectiveness in speaker verification. Interesting results show, that the lower variability of the synthesized signal can sometimes lead to significantly higher scores in comparison to those of the real speech. Some possible countermeasures against this type of attacks are discussed.

Keywords: speaker verification spoofing, speech synthesis

1. Introduction

Biometrics uses methods for recognition of humans based upon intrinsic physical or behavioural traits. Biometrics is used as a form of identity access management and control. Different areas of biometrics include physical biometrics, behavioural biometrics and medical biometrics [1, 2]. This paper is focused on the area of people identity verification from the acoustic characteristics of speech and studies the effectiveness of speaker verification systems spoofing using speech synthesis.

In contrast to most of the published research (see e.g. [3]), which most often checks the spoofing capability of the synthesizers in which the universal voice built from a huge amount of data of many speakers is adapted to a target voice using only several tenths of utterances of the target speaker, in this work we compare three synthesizers built entirely from the recordings of each particular target speaker. This approach was chosen to get a rough idea of the effectiveness of different types of the state of the art speech synthesis systems themselves in the Speaker Verification (SV) spoofing without the influence of adaptation. Therefore the same utterances included in the original voice database of the target speaker are also used for training and testing the synthetic voices. The training sets and testing sets are of course different. The settings of the synthesizers were optimized by a human expert to maximum naturalness, intelligibility and similarity to the original voice. They have not been by no means optimized to reach the maximum speaker verification scores.

2. Speaker verification

Speaker verification is a popular biometric identification technique [4] used for authenticating subjects using their biometrics, the speech signal. The method is attractive as it does not require direct contact with the subject (e.g. like iris and finger print recognition systems); it also does not require any special sensors, because microphones are now present on most portable hardware [5].

In speaker verification, a system decides whether the speaker is the same subject as he claims to be; thus the response is either true or false [6].

Mel Frequency Cepstral Coefficients (MFCC) are generally used to represent the acoustic parameters of the speech signal. In [7] the effect of multi-level wavelet decomposition as feature extraction method based on Artificial Neural Networks (ANNs) was studied. The work [8] presented an efficient Particle Swarm Optimization based optimization to enhance the performance of ANN for speaker recognition by means of optimizing ANN weights.

Approaches based on joint factor analysis (JFA) [9, 10], acoustic factor analysis (AFA) [11], i-vectors [12] and probabilistic linear discriminant analysis (PLDA) [13], have maintained outstanding performance in challenging evaluation scenarios, such as the Speaker Recognition Evaluation series developed by the National Institute of Standards and Technology (NIST) [14, 15].

The i-vector approach has risen to prominence as the de facto standard in recent speaker verification systems, due to its intrinsic capability to map an utterance to a single low-dimensional i-vector, turning a complex high-dimensional speaker recognition problem into a low-dimensional classical pattern recognition one [16, 17, 18, 19].

3. Speaker verification system used in the experiment

The SV system was created using KALDI research toolkit [20]; the i-vector approach [12] was used with PLDA scoring [21]. LibriSpeech corpus [22] was used for UBM training (2500 English speakers, 3 minutes of speech per each).

A part of the VoxForge database [23] was used as a test set. 400 speakers were chosen as target speakers, and other 1500 speakers as “impostors” (non-target). One minute of speech per speaker was used for enrolment. The utterances were taken from different recording sessions when available. The total number of 80 000 test utterances was used with the length of 2 to 10 seconds.

In the Speaker verification test each of the enrolled speakers has been tested against his own utterances (target) and against utterances of each of all the other speakers (non-target). The distance between a particular speaker model and the actual incoming utterance is expressed as **score**. As can be seen on Fig.1.a., the verification works well on the VoxForge data, as the score distributions of target and non-target speakers overlap only minimally.

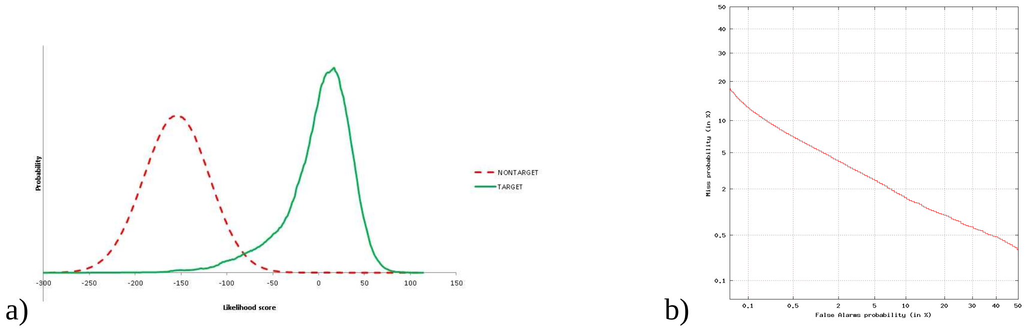

Figure 1: a) Score distributions for target and non-target speakers with the VoxForge test set; b) DET curve of the speaker verification tested on VoxForge speech database.

The reliability of speaker verification on a given test data is generally presented on the Detection Error Trade-off (DET) Curve [24]. DET-curve plotting software provided by NIST [25] was used to create the graph of the DET curve of the speaker verification task (see Fig.1.b.). The Equal Error Rate (EER) in the SV was about 3%.

Both training and testing databases include only English speech. Such a big speech databases are not available in Slovak so the authors of this paper has accomplished a number of cross-language experiments using UBM trained on English speakers for SV of Slovak speakers speaking Slovak. The effect of using different language for UBM training than for testing did not introduce such a big error that would substantially affect the results of this comparison experiment because the setting of the SV system was the same for all voices. As the results of the SV were very good when tested on several Slovak databases, it was decided to use the UBM trained on English LibriSpeech database also in this work.

4. Spoofing in speaker verification

According to [26] most biometric systems (including SV) are vulnerable to imposture. Spoofing attacks are performed on a biometric system at the sensor or acquisition level to bias score distributions toward those of genuine clients, increasing the False Acceptance Rate (FAR). [27]

Today the spoofing attacks on the SV systems are mostly realized through:

Impersonation [see e.g. 28, 29]. Impersonation refers to spoofing attacks with human-altered voices and is one of the most obvious forms of spoofing. The work [30] showed that impersonation increased FAR rates from close to 0% to between 10% and 60%.

Replay attacks [31], using speech recordings of a genuine client, or concatenation of shorter segments. The equal error rate (EER) of 1% can increase to 70% using replayed spoof attacks [32].

Voice conversion [33, 34, 35], which is a technique that electronically converts one speaker’s voice towards that of another.

Speech synthesis [36,37]. In this approach a speech synthesizer is used which is adapted to the voice of genuine clients. Using an HMM-based speech synthesiser, the FAR can rise up to 91%.

Artificial, non-speech-like tone signals [26]. The work [38] shows significant vulnerabilities to entirely artificial, non-speech-like tone signals. Certain short intervals of converted speech yield extremely high scores or likelihoods. Such intervals are not representative of intelligible speech but they are nonetheless effective in overcoming typical SV systems.

5. Speech synthesis systems used in the experiment

Three types of speech synthesizers were used in the experiment, with four voices each. The studio recordings of two male and two female speakers were used for the synthesizers training and testing. The sampling frequency was 16 kHz.

The Unit Selection synthesizer (us) was completely developed at the Institute of Informatics of the Slovak Academy of Sciences (IISAS).[39] The intonation model and model of phoneme lengths is based on CART trees. The basic concatenation elements used in this system are syllables.

The HMM synthesizer (hmm), a statistic parametric synthesizer with intonation, phoneme-lengths and spectral models, was developed using HMM-based Speech Synthesis System tools [40].

Pentaphones were used as basic elements and they were modelled using five-state HMMs. Three streams are modelled using HMMs: Duration (state durations), logarithmic fundamental frequency - logF0, and Spectral parameters (Mel Cepstral Coefficients). The speech is generated by the Mel-generalized cepstral vocoder (MGC) using a simple pulse/noise excitation. This excitation model does not fully represent natural excitation signals and generates “buzzy” speech. [41, 42].

The DNN synthesizer (dnn) was developed using Merlin toolkit for building Deep Neural Network models for statistical parametric speech synthesis [43]. It was used in combination with a front-end text processor designed at the IISAS, and the WORLD vocoder [44], version 0.2.0. WORLD decomposes input speech into three parameters: Fundamental frequency (F0), spectral envelope and aperiodicity. Representation of excitation via the band-aperiodicity function overcomes the older approach using direct excitation signal modelling [45]. The used Deep Neural network had six Feed Forward hidden layers, having 1024 hyperbolic tangent units each.

Speech data for synthesizers training and testing were obtained by studio recording of prompted Slovak utterances, accomplished by two male and two female speakers. The length of the utterances was typically one sentence, but the sets also included shorter (one or more words) and longer (up to three sentences) utterances. The female speakers *db* and *ss* recorded 8733 and 2542 utterances respectively, which is about 9 and 2.5 hours of speech. The male speakers *mr* and *sc* recorded 2486 and 1571 utterances respectively, which is about 2.5 and 1.5 hours of speech.

To illustrate the overall timbre similarity of voices we compare on Fig.1. the differences in the Long Time Average Spectra (LTAS) of the synthesized voices with respect to that of the original voice (female speaker *db*). It can be seen, that the statistical parametric synthesizers hmm and dnn exhibit peaky structure of the difference function probably caused by a well known phenomenon of their preference of the average values of pitch. In the range 600 to 6600 Hz the difference does not exceed 5 db, which suggests that spectral modeling is good. The noticeable suppression of frequencies over 7.5 kHz does not negatively affect the quality of the synthesized signal.

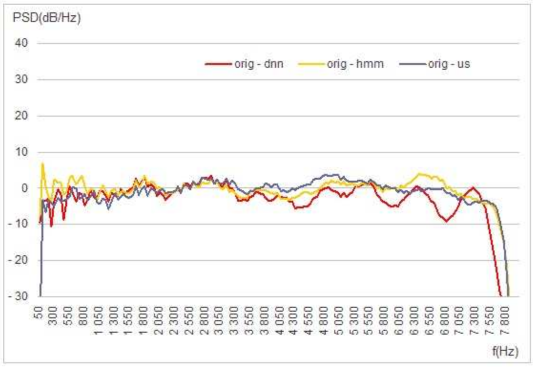

Figure 2: Differences between the LTAS of the speaker's voice and that of particular synthetic voices (female speaker *db*) . X-axis shows the frequency and y-axis gives the values of Power Spectral Density.

To get an idea on the differences in the sound of the sythesizers we present on Figure 3 spectrograms of one sentence produced by the original voice and by the three types of synthesizers. The spectrogram of the hmm synthesized utterance seems to be blurry – the formant structure is a bit less clear, the unit selection algorithm leaves observable discontinuities on the concatenation points. The dnn synthesized utterance is evidently most similar to the original voice.

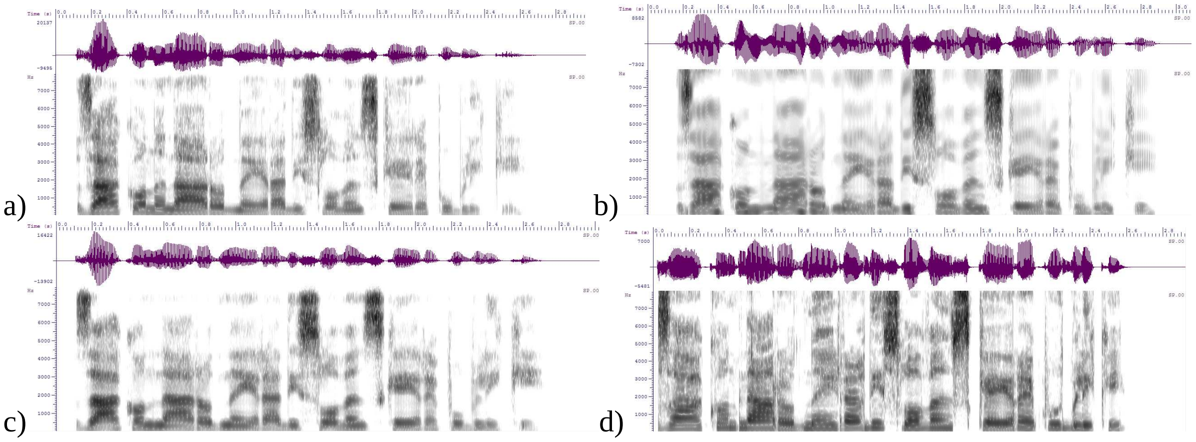

Figure 3: Spectrograms of the utterance "Zákon Národnej Rady Slovenskej Republiky" (The law of the National Council of the Slovak Republic) uttered by a) the speaker's original voice, b) HMM synthesizer - hmm, c) DNN synthesizer - dnn, and d) Unit Selection synthesizer - us (female speaker *db*).

6. Spoofing experiments

One minute of the recorded speech of each speaker was used to create their enrolments. 150 recorded utterances (same for each speaker) were chosen as test set. The text of these utterances was used as the input for the three speech synthesizers to create the spoofing utterances.

It is possible to attack the speaker verification system at various parts of the verification process. In this work the insertion of the testing and spoofing messages is done at the transmission level to avoid problems with microphone mismatch, reverberation and background noise, which will be studied later.

6.1 Results

The results of the experiments are presented in Table 1 in the form of Average PLDA SV score, its standard deviation and relative change in average PLDA SV score. The Equal Error Rates per speaker and synthesizer type are presented in Table 2.

Table 1: Average PLDA SV score, standard deviation of PLDA SV score and relative change in average PLDA SV score

	Average of score				StDev of score				Relative change of score		
Speaker	orig	us	hmm	dnn	orig	us	hmm	dnn	us	hmm	dnn
db	36.1	33.3	32.0	33.4	14.7	14.9	13.6	10.7	- 7.9%	- 11.6%	- 7.5%
mr	40.6	33.3	37.1	37.1	23.3	15.0	11.1	9.5	- 18.0%	- 8.8%	- 8.8%
sc	38.1	29.3	35.9	37.9	6.9	8.9	4.7	4.7	- 23.2%	- 5.8%	- 0.5%
ss	26.1	15.6	17.1	22.0	12.5	12.8	8.5	7.6	- 40.4%	- 34.5%	- 15.8%
Average	35.2	27.9	30.5	32.6	14.3	12.9	9.5	8.1	- 22.4%	- 15.2%	- 8.2%

Table 2: Relative change in EER

	EER[%]			
Speaker	hmm	us	dnn	Avg (hmm+us+dnn)
db	42.7%	46.3%	42.3%	43.6%
ss	31.3%	32.0%	40.0%	33.9%
sc	42.0%	28.0%	46.3%	39.3%
mr	42.3%	38.0%	40.3%	40.4%
all	44.8%	39.5%	46.9%	44,0%

Note, that the EER values are slightly bigger when computed on the data from all speakers together, as it was not possible to set a speaker-optimized threshold in this case. The value of EER equal to 50% would be reached by a classifier with a random score generator, therefore the EER in the range of about 39% to 47% is quite high, which means that spoofing is very effective.

Figure 4 depicts the probability distributions of the normalized scores reached by original speech test utterances and those reached by spoofing utterances generated by three types of synthesizers. X-axis does not represent directly the score, but it represents the PLDA score in percent normalized to the mean of the scores reached by the original utterances of the corresponding target speaker. These means are set to 100% for all of the four speakers. “m” means that the depicted function is a Gaussian model, i.e. an approximation of the distribution, and “h” are values of the histogram, i.e. the counts of the really measured values of score per 10% range. The corresponding normalized scores of all the four speakers were merged together for this measurement.

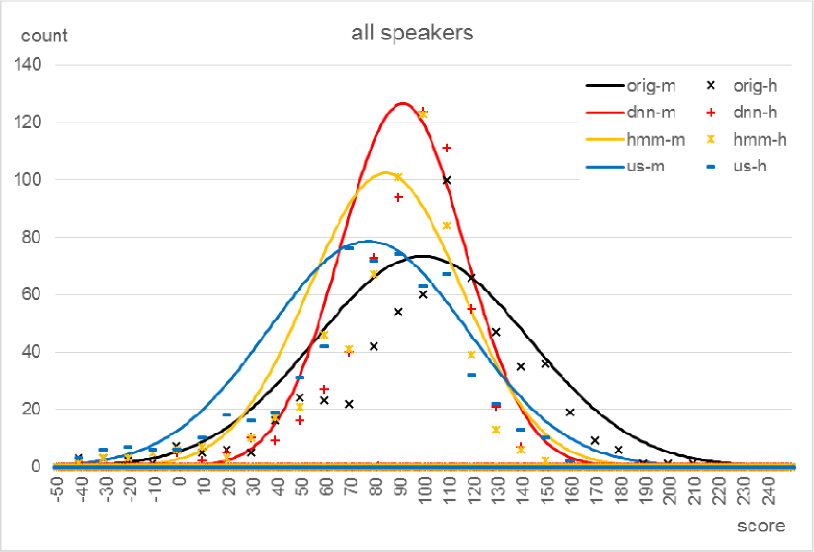

Figure 4: Probability density functions of the normalized score of original speech utterances and utterances synthesized by the three types of synthesizers. “m” means Gaussian model of the distribution and “h” means histogram-values.

7. Discussion and conclusions

As every speech utterance is different one from another, the realization and length of the enrolment greatly affect the results of the speaker verification. It was observed, that with some enrolments the PLDA score achieved by synthesized speech systematically reached higher average values than the same utterances uttered by the original voice. For these cases the enrolled speech signal probably presented deviations from the particular original voice utterances used for testing due to the speech variability, intra-speaker variability and inter-session variability. The models used in the synthesizers are obtained statistically and therefore generate speech with more general, averaged, features. This representation could have been more robust against the “atypical” enrolments, and keep consistently higher similarity to the enrolled speech samples in these cases. The statistical parametric synthesizers show considerably lower standard deviation of score than the original speech which also suggests lower variability in these voices.

If the enrolment is relatively short (1 min) and the tested utterances are also short (from one word to three sentences), as was the case in this study, it is very difficult to find effective countermeasures against the attacks using speech synthesizers. Some authors take advantage of the fact that most of the speech processing techniques neglect the phase information and they detect phase perturbations in order to detect synthetic impostors attacking SV systems. Modified Group Delay and Relative Phase Shift were used in the experiments to represent the phase information. [46] The authors of this paper are currently analysing the effectiveness of using All Pole Group Delay Features for detecting the synthesized speech attacks and will publish the results soon.

The results of the work presented in this paper can be summarized as follows: The spoofing utterances generated by all three tested types of synthesizers has reached very high PLDA scores, that would be capable of breaking protection based on the speaker verification.

For the i-vector based speaker verification system the EER has reached 39.5%, 44.8% and 46.9% for US, HMM and DNN synthesizers respectively. This also means that DNN synthesizer, which is the most up-to-date of the tested synthesis systems, has shown the highest capability of effective spoofing the Speaker Verification.

8. ACKNOWLEDGEMENT

This research was supported by the Slovak Scientific Grant Agency VEGA, grant No. 2/0197/15.

REFERENCES

1 Jucheng Yan, (2011), editor, *Biometrics*, IntechOpen Access Publisher, Croatia, ISBN: 978-953-307-618-8, [Online.] available: http://www.intechopen.com/books/biometrics

2 Pleva, M, Bours, P., Hladek, D., Juhár, J., Using current biometrics technologies for authentication in e-learning assessment, *Proceedings of the International Conference on Emerging eLearning Technologies and Applications ICETA*, Košice, (2016).

3 Wu, Z, Tomi Kinnunen, T, Evans, N., Yamagishi, J., Hanilc, C., Sahidullah, M, Sizov, A, ASVspoof 2015: the First Automatic Speaker Verification, Spoofing and Countermeasures Challenge, *Proceedings of the International Conference Interspeech* (2015).

4 Jain, A. K., Ross, A., and Prabhakar, S., An introduction to biometric recognition, *IEEE Trans. Circuits Syst. Video Technol. (Special Issue on Image- and Video-Based Biometrics)*, **vol. 14**, no. 1, (2004).

5 Fazel, A. and Chakrabartty, S., An Overview of Statistical Pattern Recognition Techniques for Speaker Verification, *IEEE Circuits and Systems Magazine*, **vol. 11**, Issue: 2, (2011).

6 Amino, K., Osanai, T., Kamada, T., Makinae, H., and Arai, T., Historical and Procedural Overview of Forensic Speaker Recognition as a Science. In: *A. Neustein and H. A. Patil, Forensic Speaker Recognition.* New York: Springer, (2012).

7 Aladwan, A. A., and Aladwan, A., A Novel Study of Biometric Speaker Identification Using Neural Networks and Multi-Level Wavelet Decomposition, *World Comput. Sci. Inf. Technol. J.*, **vol. 2**, no. 2, pp. 68–73, (2012).

8 Yadav, R., and Mandal, D., Optimization of Artificial Neural Network for Speaker Recognition using Particle Swarm Optimization, *Int. J. Soft Comput. Eng.*, **vol. 1**, no. 3, pp. 80–84, 2011.

9 Kenny, P., Joint factor analysis of speaker and session variability: theory and algorithms. *Tech. rep., CRIM.* (2005).

10 Vogt, R., Baker, B., Sridharan, S., Factor analysis subspace estimation for speaker verification with short utterances. In: *Proceedings of Interspeech 2008*, Brisbane, Australia. (2008).

11 Hasan, T. and Hansen J. H. L., Acoustic Factor Analysis for Robust Speaker Verification, *IEEE Trans. Audio. Speech. Lang. Processing*, **vol. 21**, no. 4, pp. 842–853, Apr. 2013.

12 Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., Frontend factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process.*, **vol. 19**, Issue: 4, 2010.

13 Kenny, P., Bayesian speaker verification with heavy tailed priors. In: *Proceedings of the Odyssey Speaker and Language Recogntion Workshop*, Brno, Czech Republic, (2010).

14 NIST,. The NIST year 2008 speaker recognition evaluation plan. *Tech. rep.*, NIST., [Online.] available: http://www.itl.nist.gov/iad/mig/tests/sre/ 2008/

15 NIST, 2010. *The NIST year 2010 speaker recognition evaluation plan. Tech. rep.*, NIST., [Online.] available: http://www.itl.nist.gov/iad/mig/tests/sre/ 2010/

16 Kanagasundaram, D. Dean, S. Sridharan, J. Gonzalez-Dominguez, J. Gonzalez-Rodriguez, and D. Ramos, Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques, *Speech Communnication*, **vol. 59**, pp. 69–82, Apr. 2014.

17 Kanagasundaram, A., Dean, D., Sridharan, S., McLaren, M., and Vogt, R., I-vector based speaker recognition using advanced channel compensation techniques, *Comput. Speech Lang.*, **vol. 28**, no. 1, pp. 121–140, (2014).

18 Min, J., Kua, K., Epps, J., and Ambikairajah E., i-Vector with sparse representation classification for speaker verification, *Speech Communnication*, **vol. 55**, no. 5, pp. 707–720, (2013).

19 Alam, M. J., Kinnunen, T., Kenny, P., Ouellet, P., andO'Shaughnessy, D., Multitaper MFCC and PLP features for speaker verification using i-vectors, *Speech Communnication*, **vol. 55**, no. 2, pp. 237–251, Feb. 2013.

20 Kaldi [Online.] available: http://kaldi-asr.org/, January 2016

21 Kanagasundaram, A., Vogt, R., Dean, D., Sridharan, S.: PLDA based Speaker Recognition on Short Utterances. In: *Proceedings of Oddysey Speaker and Language Recognition Workshop.* (2012)

22 LibriSpeech, [Online.] available: http://www.openslr.org/12/

23 VoxForge, [Online.] available: http://www.voxforge.org/

24 Martin, A. F. et al., "The DET Curve in Assessment of Detection Task Performance", *Proc. Eurospeech '97*, Rhodes, Greece, Vol. 4, pp. 1899–1903, (1997).

25 DET, [Online.] available: http://www.itl.nist.gov/iad/mig/tools/

26 Alegre F. et al., Spoofing countermeasures for the protection of automatic speaker recognition from attacks with artificial signals, in "INTERSPEECH 2012, *13th Annual Conference of the International Speech Communication Association*, Portland, United States (2012)

27 Evans N. et al., Spoofing and countermeasures for automatic speaker verification, in "INTERSPEECH 2013, *13th Annual Conference of the International Speech Communication Association*, Lyon, France (2013)

28 Farrs M., et al., 2008, How vulnerable are prosodic features to professional imitators?, In: *Odyssey*, 2008.

29 Blomberg M., et al., 2004 Speaker verification scores and acoustic analysis of a professional impersonator, in: *FONETIK*, 2004.

30 Lau Y., et al., Testing voice mimicry with the YOHO speaker verification corpus, in *Knowledge-Based Intelligent Information and Engineering Systems*. Springer, 2005, pp. 907–907, (2005)

31 Lindberg J. and Blomberg M., Vulnerability in speaker verification - a study of technical impostor techniques, in *European Conference on Speech Communication and Technology*, pp. 1211–1214. (1999)

32 Villalba J. and Lleida E., Speaker verification performance degradation against spoofing and tampering attacks, in *FALA workshop*, 2010, pp. 131–134. (2010)

33 Perrot P., et al.,Voice forgery using ALISP : Indexation in a Client Memory, *IEEE International Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 17 – 20. (2005)

34 Bonastre J.F., et al. Artificial impostor voice transformation effects on false acceptance rates," in *Proc. Interspeech*, pp. 2053–2056, (2007)

35 Kinnunen T., et al., Vulnerability of Speaker Verification Systems Against Voice Conversion Spoofing Attacks: the case of Telephone Speech, *Proc. ICASSP*, 2012, pp. 4401–4404, (2012)

36 Masuko T., et al., On the security of HMM-based speaker verification systems against imposture using synthetic speech, in *EUROSPEECH*, (1999)

37 De Leone P. L., et al., Revisiting the security of speaker verification systems against imposture using synthetic speech," in *IEEE International Conference on Acoustics Speech and Signal Processing*, march 2010, pp. 1798 – 1801, (2010)

38 Alegre F., et al., 2012b, On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals, *Proceedings of EUSIPCO*, (2012).

39 Darjaa S., Rusko M. and Trnka M., Three Generations of Speech Synthesis Systems in Slovakia *Proceedings of SPECOM*, St. Petersburg, Russia, 297-302, (2006).

40 HTS, *HMM-based Speech Synthesis System*. [Online.] available: http://hts.sp.nitech.ac.jp/

41 Hu, Q., Richmond, K., Yamagishi, J., Latorre, J., An experimental comparison of multiple vocoder types, *Proceedings of the 8th ISCA Speech Synthesis Workshop*, pp. 136-140, Barcelona, Spain (2013)

42 Kim, S.-J.. Hahn, M.-S., Two-band excitation for HMM-based speech synthesis, *IEICE Trans. Inf. & Syst.*, **vol. E90-D**, no.1, pp.378-381, Jan. 2007.

43 Zhizheng Wu, Oliver Watts, Simon King, "Merlin: An Open Source Neural Network Speech Synthesis System" in *Proc. 9th ISCA Speech Synthesis Workshop (SSW9)*, September 2016, Sunnyvale, CA, (2016).

44 Morise, M., Yokomori, F., and Ozawa, K., WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, *IEICE transactions on information and systems*, **vol. E99-D**, no. 7, pp. 1877-1884, (2016).

45 Morise, M., D4C, a band-aperiodicity estimator for high-quality speech synthesis, In: *Speech Communication*, **vol. 84**, pp. 57-65, Nov. 2016.

46 Saratxaga, I., Sanchez, J., Wu, Z., Hernaez,I., Navas, E., Synthetic speech detection using phase information, *Speech Communication*, **vol. 81** Issue C, Elsevier Science Publishers B. V. Amsterdam, The Netherlands, The Netherlands pp. 30-41, (2016).