On the Vulnerability of Speaker Verification to Realistic Voice Spoofing

Serife Kucur Ergünay, Elie Khoury, Alexandros Lazaridis, Sébastien Marcel
Idiap Research Institute
Centre du Parc, Martigny, Switzerland

Abstract

Automatic speaker verification (ASV) systems are subject to various kinds of malicious attacks. Replay, voice conversion and speech synthesis attacks drastically degrade the performance of a standard ASV system by increasing its false acceptance rates. This issue raised a high level of interest in the speech research community where the possible voice spoofing attacks and their related countermeasures have been investigated. However, much less effort has been devoted in creating realistic and diverse spoofing attack databases that foster researchers to correctly evaluate their countermeasures against attacks. The existing studies are not complete in terms of types of attacks, and often difficult to reproduce because of unavailability of public databases. In this paper we introduce the voice spoofing data-set of ***AVspoof****, a public audio-visual spoofing database. AVspoof includes ten realistic spoofing threats generated using replay, speech synthesis and voice conversion. In addition, we provide a set of experimental results that show the effect of such attacks on current state-of-the-art ASV systems.*

Index Terms: spoofing, countermeasure, replay attack, speech synthesis, voice conversion, speaker verification

1. Introduction

Over the last years, a lot of progress has been seen in the study of face and voice biometrics. With the successful use of deep neural networks [26] and i-vectors [6, 17] that take advantage of the increasing data volume, current state-of-the-art face and speaker recognition systems are able to effectively deal with intra-class and inter-class variability. As a matter of fact, many automatic face recognition or speaker verification systems are made available in operational use for a wide range of applications. However, the prevalent usage of those systems raised some security concerns, since a copy or an artificial sample of face/voice biometrics can easily be reproduced to "fool" the identity verification systems.

A number of literature studies in automatic face [4, 11] and speaker [2, 30, 27] verification showed the vulnerability of the state-of-the-art systems to the spoofing attacks. These findings led to the increase of interest on the spoofing topic in the biometrics research community. Moreover, anti-spoofing challenges for face [5], and voice[1] biometrics are launched to promote more interest in the spoofing/anti-spoofing research.

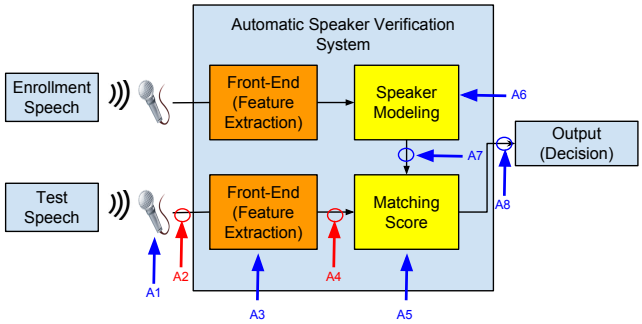

Figure 1: Standard ASV system and eight possible weak links. Attacks could be [1]: A1) Fake biometric, A2) replay attacks, A3) override feature extractor , A4) synthetic features/samples, A5) override matcher , A6) changing model, A7) intervention to the channel, A8) override output.

The findings in (anti-)spoofing research, including the countermeasure techniques are however based on *in-house* spoofing attacks. As a result, they are prone to be biased and need to be tested in a standard and comparative way. For researchers to fairly and "quickly" evaluate their systems towards spoofing attacks and to propose more generalized countermeasures, a spoofing database covering a large number of attacks is of key importance. Recently, efforts have been made to create such databases where the attacks targeting face and speaker verification systems are considered. In this study, we focus solely on spoofing automatic speaker verification systems.

To the best of our knowledge there are only few existing speech spoofing databases. In [31, 33], voice conversion and replay attacks are generated using the RSR2015 database [19]. In [32], a number of speech synthesis and

[1] http://www.spoofingchallenge.org

voice conversion techniques are implemented. In [18, 25], attacks based on various voice conversion methods are generated. However, none of them includes simultaneously all the types of attacks, namely replay, speech synthesis and voice conversion. Another drawback of these databases is the fact that they only provide the logical access (i.e. feature level, type A4) simulations and do not comprise the physical access (i.e. sensor level, type A2) case, which constitutes more realistic scenarios (see Fig. 1). Furthermore, the lack of different recording sessions can be observed in some of these databases, or the lack of adequate number of subjects can be observed in others. These are the main reasons that hinder the generalization of both spoofing and anti-spoofing techniques.

In this paper, we introduce a unique spoofing database that includes all the types of spoofing attacks. This database is expected to constitute a reliable benchmark for voice spoofing techniques and countermeasures, since it includes wide diversity in terms of sessions, environmental setups and acquisition devices, and various spoofing scenarios within challenging conditions. Additionally, we provide the physical access simulations, i.e. the replayed form of the speech synthesis and the voice conversion attacks, which brings another novelty compared to state-of-the-art work.

The reminder of this paper is organized as follows. Section 2 briefly presents the state-of-the-art ASV systems. Section 3 describes the data and methods used for creating the attacks. Section 4 shows the experimental results for ASV systems with and without countermeasures. Finally, Section 5 concludes the paper.

2. Automatic Speaker Verification

Automatic speaker verification is regularly evaluated by the National Institute of Standards and Technology (NIST)[2] since 1996 in the context of the NIST speaker recognition evaluation (SRE) series. During this series, many techniques have been proposed such as *Gaussian mixture models* (GMM) [24], *inter-session variability* (ISV) modeling [28], *joint factor analysis* (JFA) [15], or recently, *i-vectors* [7]. One common thread with current successful state-of-the-art approaches is their ability to cope with session variability that mainly comes from acoustic environments and communication channels.

In this study, two of the above techniques are evaluated in terms of spoofing vulnerability: ISV and i-vectors.

ISV aims to estimate session variation like noise in order to compensate for it. It is assumed that session variability results in an additive offset to the mean super-vector $\boldsymbol{g}_i$ of the client model. Given the j-th utterance $\boldsymbol{\mathcal{O}}_{i,j}$ of a speaker i the mean super-vector $\boldsymbol{\mu}_{i,j}$ of the GMM is:

$$\boldsymbol{\mu}_{i,j} = \boldsymbol{m} + \boldsymbol{U}\boldsymbol{x}_{i,j} + \boldsymbol{D}\boldsymbol{z}_i \,, \tag{1}$$

[2] http://www.nist.gov/itl/iad/mig/sre.cfm

(a) Microphone (b) Phone1 (c) Phone2

Figure 2: Devices used for speech data acquisition.

where $\boldsymbol{m}$ is the GMM mean super-vector of the universal background model (UBM), $\boldsymbol{U}$ is a subspace that constrains the possible session effects, $\boldsymbol{x}_{i,j}$ is its associated latent session variable ($\boldsymbol{x}_{i,j} \sim \mathcal{N}(0, \boldsymbol{I})$), while $\boldsymbol{D}$ and $\boldsymbol{z}_i$ represent the client-specific offset.

The total variability modeling aims at extracting a low-dimensional factor $\boldsymbol{w}_{i,j}$, called "i-vector". It relies on the definition of a total variability subspace $\boldsymbol{T}$ and can be described by:

$$\boldsymbol{\mu}_{i,j} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}_{i,j} \,, \tag{2}$$

The i-vector approach only acts as a front-end extractor and does not perform session compensation or scoring. Thus, several techniques are commonly applied to i-vectors such as *whitening* [10], *length-normalization* [10], linear discriminant analysis (LDA) [9], within-class convariance normalization (WCCN) [12], *probabilistic linear discriminant analysis* (PLDA) [23].

3. Realistic Voice Spoofing Attacks

3.1. Common Set-up

The AVspoof database [3] is intended to provide stable, non-biased spoofing attacks in order for researchers to test both their ASV systems and anti-spoofing algorithms. The attacks are created based on newly acquired audio recordings. The data acquisition process lasted approximately two months with 44 persons, each participating in several sessions configured in different environmental conditions and setups. After the collection of the data, the attacks, more precisely, replay, voice conversion and speech synthesis attacks were generated.

The data acquisition process is divided into four different sessions, each scheduled several days apart in different setups and environmental conditions (e.g. different in terms of background noise, reverberation, etc.) for each of 31 male and 13 female participants. The first session which is supposed to be used as training set while creating the attacks, was performed in the most controlled conditions. Besides, the conditions for the last three sessions dedicated to test trials were more relaxed in order to grasp the challenging

[3] https://www.idiap.ch/dataset/avspoof

	Session 1	Session 2-4	Total
read	40 sentences	10 sentences	25.96 hours
pass	5 pass-phrases	5 pass-phrases	4.73 hours
free	$\geq$ 5 min.	$\geq$ 3 min.	38.51 hours

Table 1: The statistics of the collected data in terms of session, recording type and acquisition device.

scenarios. The audio data were recorded by three different devices (Fig. 2) including (a) one good-quality microphone, AT2020USB+, and two mobiles, (b) Samsung Galaxy S4 (phone1) and (c) Iphone 3GS (phone2) .

The positioning of the devices was stabilized for each session and each participant in order to standardize the recording settings.

For each session, the participant was subjected to three different data acquisition protocols as in the following:

- Reading part (*read*): 10/40 pre-defined sentences are read by the participant.
- Pass-phrases part (*pass*): 5 short prompts are read by the participant.
- Free speech part (*free*): The participant speaks freely about any topic for 3 to 10 minutes.

The number, the length, as well as the content of the sentences for the reading and pass-phrases part are carefully selected in order to satisfy the constraints in terms of readability, data acquisition and attack quality. Similarly, the minimum duration of the free speech part is also determined according to our preliminary investigations mostly on the voice conversion attacks for which the free speech data would be included in the training set. Please refer to Table 1 for the statistics of the collected data.

3.2. Attacks

In the spoofing attack creation phase, we considered creating spoofing trials for the text-dependent utterances of the testing data, i.e. reading parts of sessions 2-4 and the pass-phrases of all four sessions. As a preliminary step before the creation of the attacks, the speech data originally recorded at 44.1 KHz sampling rate is down-sampled to 16 KHz.

There are four main spoofing attacks for ASV systems: Impersonation, replay, speech synthesis and voice conversion [8]. As the impersonation is known not to be a serious threat for ASV systems [8], we did not include it in our database. For the remaining three spoofing types, we designed ten different scenarios (see Table 2). We gave special attention to physical access attacks. These attacks are more realistic than logic access attacks considering the fact that the attacker often has no direct access to the system. The acquisition devices (sensors) are open to anyone, therefore more subjected to such attacks.

Attacks	**Num. of trials per speaker**		**Total num. of trials**	
	Male	*Female*	*Male*	*Female*
Replay-phone1	50	50	1,550	650
Replay-phone2	50	50	1,550	650
Replay-laptop	50	50	1,550	650
Replay-laptop-HQ	50	50	1,550	650
Speech-Synthesis-LA	35	35	1,085	455
Speech-Synthesis-PA	35	35	1,085	455
Speech-Synthesis-PA-HQ	35	35	1,085	455
Voice-Conversion-LA	1,500	600	46,500	7,800
Voice-Conversion-PA	1,500	600	46,500	7,800
Voice-Conversion-PA-HQ	1,500	600	46,500	7,800

Table 2: Number of spoofing trials per gender.

3.2.1 Replay Attacks

A replay attack consists of replaying a pre-recorded speech to an ASV system. We assume that the ASV system has a good quality microphone and the replay attack targets this sensor. Three different scenarios are considered:

- *Replay-phone1*: Replay attack using the data captured by the Samsung mobile. The speech recorded by this mobile is replayed using its own speakers and re-recorded by the microphone of the ASV system.
- *Replay-phone2*: Replay attack using the data captured by the iPhone mobile. The speech recorded by this mobile is replayed using its own speakers and re-recorded by the microphone of the ASV system.
- *Replay-laptop*: Replay attack using the data captured by the microphone of the ASV system. The speech recorded by this microphone is replayed using the laptop speakers and re-recorded again by the microphone of the system.
- *Replay-laptop-HQ*: Replay attack using the data captured by the microphone of the ASV system. The speech recorded by this microphone is replayed using external high-quality loudspeakers and re-recorded using the microphone of the ASV system.

3.2.2 Speech Synthesis Attacks

The speech synthesis attacks were based on statistical parametric speech synthesis (SPSS) [38]. More specific, hidden Markov model (HMM)-based speech synthesis technique [36] was used. Today, HMM-based speech synthesis produces very high quality synthetic speech, sometimes achieving even higher performance than unit-selection concatenation synthesis [38]. In this approach, a unified HMM framework is used to simultaneously model the spectrum,

fundamental frequency and segment duration, creating a mapping between the input (linguistic) features and the output (acoustic) features. Additionally HMM-based synthesis with the use of adaptation and interpolation techniques [35], [22], offers flexibility in changing speaker's and voice characteristics, speaking style, emotions, etc.

In this paper, the HTS version 2.1 toolkit [13] was used for building the HMM models. Specifically, the implementation from the EMIME project [29] was taken. Five-state, left-to-right, no-skip HSMMs (hidden semi-Markov models) were used [39]. A standard, in HMM-based speech synthesis, set of input features was used, composed of phonetic and prosodic features such as phone identity, identity of the two previous and next phones, number of syllables in a word, accented/stressed syllable, etc., adding to a 53 feature set [37]. The speech parameters which were used for training the HSMMs were 39 order mel-cepstral coefficients, log-F0 and 21-band aperiodicities, along with their first and second derivatives, extracted every 5 ms. An average voice model was first built using SAT training [3] on the SI-84 (si-tr-s) subset of the American English multi-speakers Wall Street Journal (WSJ) corpora [20], containing of approximately $7,100$ sentences. The average voice model was then adapted to each speaker, using 40 sentences (i.e. Session 1-read) of the respective speaker. For the adaptation, the constrained structural maximum a posteriori linear regression (CSMAPLR) [34] approach was used. STRAIGHT [14] was used for the analysis and synthesis phase of the HMM-based speech synthesis.

Accordingly, three scenarios were involved:

- *Speech-Synthesis-LA*: Speech synthesis via logical access. The synthesized speech is directly presented to the ASV system without being re-recorded.
- *Speech-Synthesis-PA*: Speech synthesis via physical access. The synthesized speech is replayed using the laptop speakers and re-recorded by the microphone of the ASV system.
- *Speech-Synthesis-PA-HQ*: Speech synthesis via high-quality physical access. The synthesized speech is replayed using external high-quality loudspeakers and re-recorded by the microphone of the ASV system.

3.2.3 Voice Conversion Attacks

The voice conversion attacks were created using Festvox[4]. A conversion function for each pair of source-target speaker is found based on the learned GMM model/parameters by using the source and target speakers training data. We did not consider cross-gender voice conversion attacks, that is only male-to-male and female-to-female conversions were taken into account. As in the case of speech synthesis, three possible scenarios are involved:

- *Voice-Conversion-LA*: Voice conversion via logical access. The converted speech is directly presented to the system without being re-recorded.
- *Voice-Conversion-PA*: Voice conversion via physical access. The converted speech is replayed using the speakers of the laptop and re-recorded by the microphone of the ASV system.
- *Voice-Conversion-PA-HQ*: Voice conversion via high-quality physical access. The converted speech is replayed using external high-quality loudspeakers and re-recorded by the microphone of the ASV system.

4. Experimental Results

In this section, we evaluate the vulnerability of the ASV systems introduced in Section 2 under the various types of spoofing attacks. In all our experiments[5] we use Spear [16], an open-source speaker recognition tool based on Bob[6]. For both ISV and I-Vector systems, the same voice activity detection (VAD), feature extraction and UBM training steps are performed. For VAD, the modulation of the energy around 4Hz is applied. A 60-dimensional feature vector is constructed for each speech utterance. This vector includes 19 MFCCs and the energy, with their first and second derivatives. The UBM parameters are estimated using the MOBIO database [21], with 256 Gaussian components. The subspace dimension of the ISV system is set to 50. As for TV system, the dimension of i-vectors is set to 100 while the rank of LDA and the dimensions of the intra and inter speaker covariances of the PLDA model are all set to 50.

Table 3 illustrates the results of the ISV and TV systems without spoofing scenarios. It shows that the ISV results are better than TV results on both Male and Female trials. This might be due to the relatively small amount of training data used to train the subspace models of the TV system.

Table 4 shows the *spoofing false acceptance rates* (SFAR) of both ISV and TV systems with regards to each of the spoofing attacks. It is clear that all attacks give rise to high SFARs for both ISV and TV systems. More specifically, the speech synthesis and voice conversion attacks seem to be very effective in spoofing the ASV system. For instance, the speech synthesis presented directly (i.e. *logical access*) to the TV-based system can produce a SFAR of 96.5% on Male trials and 81.5% for Female trials.

It is also worth noting that the ASV system is vulnerable to physical access attacks although the SFAR rates are lower than those of logical access attacks. For example, in

[4] http://festvox.org

[5] This study is reproducible using this package: https://pypi.python.org/pypi/xspear.btas2015

[6] http://www.idiap.ch/software/bob/

	Male		Female	
	ISV	*TV*	*ISV*	*TV*
Threshold	*0.597*	*43.97*	*0.690*	*44.63*
EER (%)	4.9	6.9	10.6	17.5

Table 3: Performance of the ISV and TV systems with natural speech (i.e. without spoofing attacks).

Attack type	**Male**		**Female**	
	ISV	*TV*	*ISV*	*TV*
Replay-phone1	19.2	29.1	12.2	11.8
Replay-phone2	45.9	27.7	23.1	11.1
Replay-laptop	45.3	39.8	35.7	32.2
Replay-laptop-HQ	74.1	77.4	68.5	69.4
Speech-Synthesis-LA	97.0	96.5	83.5	81.5
Speech-Synthesis-PA	65.9	60.6	67.9	69.5
Speech-Synthesis-PA-HQ	94.1	93.5	83.7	83.7
Voice-Conversion-LA	93.4	92.6	71.2	71.6
Voice-Conversion-PA	77.4	84.0	50.7	75.8
Voice-Conversion-PA-HQ	89.3	88.8	73.0	73.0

Table 4: Verification performance of the ISV and the TV systems under spoofing attacks. SFAR (%) are computed using the thresholds given in Table 3.

the case of *Speech-Synthesis-PA*, that is, speech synthesis attacks replayed with normal laptop speakers, the SFAR is 60.6% instead of 96.5% on Male trials for a TV-based system. However, interestingly, the physical access attacks that are replayed with high-quality loudspeakers have spoofing rates close to the logical access attacks. For example, this is the case for *Speech-Synthesis-PA-HQ* with a SFAR of 93.5% for a TV-based system on Male trials.

5. Conclusions

In this paper we provide an experimental study on the vulnerability of ASV systems to realistic spoofing attacks. We introduce AVspoof, a public audio-visual spoofing database that includes ten realistic voice spoofing threats generated using replay, speech synthesis and voice conversion attacks. We also provide a set of experimental results that show the effect of these spoofing attacks on two state-of-the-art ASV systems. Future work will mainly focus on generalized countermeasure that can effectively deal with the various types of attacks presented in this study.

6. Acknowledgment

The development leading to this toolbox has received funding from the Swiss National Science Foundation (SNSF) under the LOBI project and from the European Community's Seventh Framework Programme (FP7) under grant agreement 284989 (BEAT) and the Swiss Center for Biometrics Research and Testing.

References

[1] Z. Aktar. *Security of Multimodal Biometric Systems against Spoofing Attacks*. PhD thesis, University of Cagliari, 2012.

[2] F. Alegre, A. Amehraye, and N. W. D. Evans. Spoofing countermeasures to protect automatic speaker verification from voice conversion. In *ICASSP 2013, 38th IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013.

[3] T. Anastasakos, J. McDonough, and J. Makhoul. Speaker adaptive training: a maximum likelihood approach to speaker normalization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 1043–1046 vol.2, Apr 1997.

[4] A. Anjos and S. Marcel. Counter-measures to photo attacks in face recognition: A public database and a baseline. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–7, Oct 2011.

[5] M. Chakka, A. Anjos, S. Marcel, R. Tronci, D. Muntoni, G. Fadda, M. Pili, N. Sirena, G. Murgia, M. Ristori, F. Roli, J. Yan, D. Yi, Z. Lei, Z. Zhang, S. Li, W. Schwartz, A. Rocha, H. Pedrini, J. Lorenzo-Navarro, M. Castrillon-Santana, J. Maatta, A. Hadid, and M. Pietikainen. Competition on counter measures to 2-d facial spoofing attacks. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–6, Oct 2011.

[6] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):788–798, 2011.

[7] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Trans. on Audio, Speech, and Language Processing*, 2011.

[8] N. Evans, T. Kinnunen, and J. Yamagishi. Spoofing and countermeasures for automatic speaker verification. In *INTERSPEECH*, pages 925–929, 2013.

[9] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 19:179–188, 1936.

[10] D. Garcia-Romero and C. Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems. In *Interspeech*, pages 249–252, 2011.

[11] A. Hadid. Face biometrics under spoofing attacks: Vulnerabilities, countermeasures, open issues, and research directions. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 113–118, June 2014.

[12] A. Hatch, S. Kajarekar, and A. Stolcke. Within-class covariance normalization for SVM-based speaker recognition. In *9th Intl. Conf. on Spoken Language Processing (ICSLP)*, 2006.

[13] HTS. HMM-based speech synthesis system version 2.1. 2010.

[14] H. Kawahara, J. Estill, and O. Fujimura. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. In *MAVEBA*, 2001.

[15] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Joint factor analysis versus eigenchannels in speaker recognition.

IEEE Trans. on Audio, Speech, and Language Processing, 15(4):1435–1447, May 2007.

[16] E. Khoury, L. El Shafey, and S. Marcel. Spear: An open source toolbox for speaker recognition based on Bob. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[17] E. Khoury, L. El Shafey, C. McCool, M. Günther, and S. Marcel. Bi-modal biometric authentication on mobile phones in challenging conditions. *Image and Vision Computing*, 2014.

[18] E. Khoury, T. Kinnunen, A. Sizov, Z. Wu, and S. Marcel. Introducing i-vectors for joint anti-spoofing and speaker verification. In *Proc. Interspeech*, 2014.

[19] A. Larcher, K.-A. Lee, B. Ma, and H. Li. Rsr2015: Database for text-dependent speaker verification using multiple pass-phrases. In *INTERSPEECH*, 2012.

[20] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti. The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 357–362, San Juan, US, November 2005.

[21] C. McCool, S. Marcel, A. Hadid, M. Pietikainen, P. Matejka, J. Cernocky, N. Poh, J. Kittler, A. Larcher, C. Levy, D. Matrouf, J.-F. Bonastre, P. Tresadern, and T. Cootes. Bi-modal person recognition on a mobile phone: using mobile phone data. In *IEEE ICME Workshop on Hot Topics in Mobile Multimedia*, 2012.

[22] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi. A style control technique for HMM-based expressive speech synthesis. *IEICE - Trans. Inf. Syst.*, E90-D(9):1406 – 1413, 2007.

[23] S. J. Prince and J. H. Elder. Probabilistic linear discriminant analysis for inferences about identity. In *IEEE International Conference on Computer Vision (ICCV)*, volume 0, pages 1–8, 2007.

[24] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 2000.

[25] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel. Joint speaker verification and anti-spoofing in the i-vector space. *IEEE Transactions on Information Forensics and Security (under revision)*, 2015.

[26] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1701–1708, June 2014.

[27] J. Villalba and E. Lleida. Speaker verification performance degradation against spoofing and tampering attacks. In *FALA workshop*, pages 131–134, 2010.

[28] R. Vogt and S. Sridharan. Explicit modelling of session variability for speaker verification. *Computer Speech & Language*, 22(1):17–38, 2008.

[29] M. Wester, J. Dines, M. Gibson, H. Liang, Y.-J. Wu, L. Saheer, S. King, K. Oura, P. N. Garner, W. Byrne, Y. Guan, T. Hirsimäki, R. Karhila, M. Kurimo, M. Shannon, S. Shiota, J. Tian, K. Tokuda, and J. Yamagishi. Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project. In *SSW7*, pages 192–197, 2010.

[30] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li. Spoofing and countermeasures for speaker verification: A survey. *Speech Communication*, 66(0):130 – 153, 2015.

[31] Z. Wu, S. Gao, E. S. Cling, and H. Li. A study on replay attack and anti-spoofing for text-dependent speaker verification. In *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*, pages 1–5, Dec 2014.

[32] Z. Wu, A. Khodabakhsh, C. Demiroglu, J. Yamagishi, D. Saito, T. Toda, and S. King. Sas: A speaker verification spoofing database containing diverse attacks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[33] Z. Wu, A. Larcher, K.-A. Lee, E. Chng, T. Kinnunen, and H. Li. Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints. In *INTERSPEECH*, pages 950–954, 2013.

[34] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai. Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(1):66–83, Jan 2009.

[35] J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano. Model adaptation approach to speech synthesis with diverse voices and styles. In *ICASSP*, pages 1233–1236, 2007.

[36] T. Yoshimura, K. Tokuda, T. Kobayashi, T. Masuko, and T. Kitamura. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *EUROSPEECH*, 1999.

[37] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda. The HMM-based speech synthesis system (HTS) version 2.0. In *Proceedings of the 6th ISCA Speech Synthesis Workshop*, pages 294–299, 2007.

[38] H. Zen, K. Tokuda, and A. W. Black. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064, 2009.

[39] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Hidden semi-Markov model based speech synthesis. In *Proc. of ICSLP*, 2004.