IDIAP RESEARCH REPORT

JOINT OPERATION OF VOICE BIOMETRICS AND PRESENTATION ATTACK DETECTION

Pavel Korshunov Sébastien Marcel

Idiap-RR-25-2016

OCTOBER 2016

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Joint Operation of Voice Biometrics and Presentation Attack Detection

Pavel Korshunov and Sébastien Marcel
Biometrics group, Idiap Research Institute,
Martigny, Switzerland
{pavel.korshunov,sebastien.marcel}@idiap.ch

Abstract

Research in the area of automatic speaker verification (ASV) has advanced enough for the industry to start using ASV systems in practical applications. However, as it was also shown for fingerprints, face, and other verification systems, ASV systems are highly vulnerable to spoofing or presentation attacks, limiting their wide practical deployment. Therefore, to protect against such attacks, effective anti-spoofing detection techniques, more formally known as presentation attack detection (PAD) systems, need to be developed. These techniques should be then seamlessly integrated into existing ASV systems for practical all-in-one solutions. In this paper, we focus on the integration of PAD and ASV systems. We consider the state of the art i-vector and ISV-based ASV systems and demonstrate the effect of score-based integration with a PAD system on the verification and attack detection accuracies. In our experiments, we rely on AVspoof database that contains realistic presentation attacks, which are considered by the industry to be the threat to practical ASV systems. Experimental results show a significantly increased resistance of the joint ASV-PAD system to the attacks at the expense of slightly degraded performance for scenarios without spoofing attacks. Also, an important contribution of the paper is an open source and an online-based implementations of the separate and joint ASV-PAD systems.

1. Introduction

Similar to fingerprint sensors and face recognition systems, automatic speaker verification (ASV) systems were shown to be highly vulnerable to spoofing or presentation attacks [9]. The ease with which an ASV system can be spoofed motivated researchers into developing anti-spoofing detection mechanisms, i.e., presentation attack detection (PAD) systems, that can accurately and efficiently distinguish between genuine speech and presentation attacks. Several approaches have been proposed recently, mostly focusing on the feature extraction component of

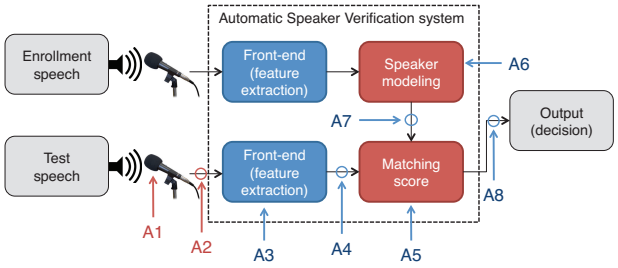


Figure 1: Possible attacks places in a typical ASV system.

PAD systems. A survey by Wu *et al.* [14] provides a comprehensive overview of the presentation attacks and the currently available PAD methods. Most of the methods use features based on the audio spectrogram in combination with a classifier, e.g., based on Gaussian mixture models (GMM), or a learning network.

However, developing a presentation attack detection method is not enough for practical use. Such PAD system should be seamlessly and effectively integrated with an existing ASV system. The goal for the resulted joint system is to be resistant to presentation attacks just like its PAD component and, in the same time, have the same verification accuracy as its ASV component.

In this paper, we integrate speaker verification and presentation attack detection systems via learning-based score fusion, as per approach originally proposed in [2] for face recognition systems. This approach allows to separate genuine data of the valid users, who are trying to be verified by the system, from both presentation attacks and genuine data of the non-valid users or so-called *zero-impostors*. For an attack detection component of the joint ASV-PAD system, we consider two different approaches: an extension of the LBP-based method [1], which uses histograms of local binary patterns (LBP) computed from an audio spectrogram as features for a logistic regression classifier, and an approach that uses MFCC features with GMM-based classifier. For ASV system, we adopt verification approaches based on ISV modeling [13] and *i-vectors* [5].

To demonstrate the feasibility of this integration approach in the context of speech biometrics and to ensure reproducibility of the results, we have implemented PAD, ASV, and joint ASV-PAD systems as open source. We provide two types of implementations: a more traditional stand-alone downloadable package[1] (LBP-based PAD, *i-vector* based ASV, and joint ASV-PAD systems), for which database need to be provided in order to compute results, and a web-based implementation (LBP- and MFCC-based PAD, ISV-based ASV, and joint systems) based on the open source BEAT platform[2], which, inline with the recent trend of open source web-based platforms such as Google's TensorFlow[3], allows to run, evaluate, and create biometric experiments using several comprehensive databases provided internally in BEAT.

Given the complexity of a practical ASV system, several different modules of the system are prone to attacks, as illustrated by Figure 1. Depending on the usage scenario, two of the most vulnerable places for spoofing attacks in an ASV system are marked by 'A1' (aka 'physical access' as defined in [15] or presentation attacks) and 'A2' (aka 'logical access' attacks as defined in [15]) in the figure. Since generating 'logical access' attacks is relatively easier, initially available databases and first research results focused mostly on this type of attacks.

In this paper, we focus on presentation attacks, because they are more realistic in practical scenarios and are considered to be a serious threat by the industry, as it is reflected in the ISO standard DIS 30107-1 [6]. Presentation attacks assume that either a stolen set of user's samples or an automatically generated set of samples is replayed to a microphone of an attacked ASV system with an attempt to mimic the genuine registered user. AVspoof[5] database [9] is the first challenging database that contains a comprehensive set of presentation attacks, including, (i) the direct replay attacks when a genuine data is played back using a laptop and two phones (Samsung Galaxy S4 and iPhone 3G), (ii) synthesized speech replayed with a laptop, and (iii) an attack data, generated using a voice conversion algorithm, replayed with a laptop.

Therefore, this paper has the following main contributions:

- Integration of PAD and ASV systems into one joint ASV-PAD system based on score fusion;
- Open source implementations of two PAD and two ASV state of the art systems;
- An extensive and reproducible evaluation of the considered systems on the realistic presentation attacks of AVspoof database.

[1] Source code: https://pypi.python.org/pypi/bob.paper.btas_j2016
[2] https://www.beat-eu.org/platform/
[3] https://www.tensorflow.org/

2. Related work

The research on presentation attack detection is far from being matured, especially, if compared to the significant advances in speech analysis and speaker verification. Most of the available work in speech anti-spoofing focuses on synthetic attacks, such as voice conversion, speech synthesis, and artificial signals [14], which are assumed to be fed into a verification system directly bypassing its microphone (indicated by 'A2' in Figure 1), hence, they are coined by the authors as 'logical access' attacks [15]. This type of attacks constitute ASVspoof[4] database, which was made recently available for Interspeech anti-spoofing challenge [15]. The most practical *replay attacks* (indicated by 'A1' in Figure 1), which are formally defined as *presentation attacks* by ISO standardization committee [6], received considerably less attention, since, until now, there was no dataset with such attacks. That is why AVspoof[5] is of great interest, because it is the first database that contains several types of replay attacks.

A survey by Wu *et al.* [14] provides a comprehensive overview of the spoofing attacks and the currently available attack detection methods. These methods use features mostly based on the audio spectrogram, such as spectral- and cepstral-based features [12], phase-based features [4], the combination of amplitude and phase features [10], and audio quality based features [7]. Also, a higher computational layer can be added, for instance, Alegre *et al.* [1] proposed to use histograms of Local Binary Patterns (LBP), which can be computed directly from a set of pre-selected spectral, phase-based, or other features. Most of these features are used successfully in speaker verification systems already, so, naturally, they are first to be proposed for PAD systems as well.

Besides determining 'good features for detecting presentation attacks', it is also important to correctly classify the computed feature vectors as belonging to real or spoofed data. Choosing a reliable classifier is especially important given a possibly unpredictable nature of attacks in a practical system, since it is unknown what kind of attack the perpetrator may use when spoofing the verification system. The most common approach to classification is to use one of the well-known classifiers, which is usually pre-trained on the examples of both real and spoofed data.

Different methods use different classifiers but the most common choices include logistic regression, support vector machine (SVM), and Gaussian mixture model (GMM) classifiers. The benchmarking study on 'logical access' attacks [11] finds GMMs to be more successful compared to two-class SVM (combined with an LBP-based feature extraction from [1]) in detecting synthetic spoofing attacks.

[4] http://datashare.is.ed.ac.uk/handle/10283/853
[5] https://www.idiap.ch/dataset/avspoof

Figure 2: AVspoof database recording setup.

In this paper, we adopt versions of two popular approaches for presentation attack detection: MFCC based features with two GMM-based classifiers (one for real data and one for attacks) and LBP-histograms computed on MFCC features with logistic regression classifier.

The research on automatic speaker verification is more established with regular competitions conducted by National Institute of Standards and Technology (NIST) since 1996[6]. Many techniques have been proposed with the most notable systems based on GMM, inter-session variability (ISV) modeling [13], joint factor analysis (JFA) [8], and *i-vectors* [5].

In this paper, we consider ASV systems based on ISV (BEAT-based implementation[2]) and *i-vectors* (stand-alone implementation[1]).

3. Experimental setup

One of the major factor constraining the development of effective presentation attack detection methods for speech is the lack of standard databases with a set of real (genuine) speech samples and a large variety of presentation attacks. To our knowledge, the most comprehensive database containing spoofing attacks is AVspoof[5] [9]. In this section, we describe this dataset and discuss how it is used to evaluate ASV and PAD systems.

3.1. Presentation attack database

AVspoof database contains real (genuine) speech samples from 44 participants (31 males and 13 females) recorded over the period of two months in four sessions, each scheduled several days apart in different setups and environmental conditions such as background noises. The first session was recorded in the most controlled conditions. Speech samples were recorded using three devices: laptop using microphone AT2020USB+, Samsung Galaxy S4 phone, and iPhone 3GS (see recording setup in Figure 2). The following type of recordings were made:

- Reading part: 10 or 40 pre-defined sentences are read by participants.
- Pass-phrases part: 5 short prompts are read by participants.
- Free speech part: participants speak freely about any topic for 3 to 10 minutes.

When generating presentation attacks, the assumption is that a verification system is installed on a laptop (with an internal built-in microphone) and an attacker is trying to gain access to this system by playing back to it a pre-recorded genuine data or an automatically generated synthetic data using some playback device. In AVspoof database, presentation attacks consist of (i) direct replay attacks when a genuine data is played back using a laptop with internal speakers, a laptop with external high quality speakers, Samsung Galaxy S4 phone, and iPhone 3G, (ii) synthesized speech replayed with a laptop, and (iii) converted voice attacks replayed with a laptop.

The data in AVspoof database is split into three non-overlapping subsets: training (real and spoofed samples from 4 male and 10 female participants), development or *Dev* (real and spoofed samples from 4 male and 10 female participants), and test or *Test* (real and spoofed samples from 5 male and 11 female participants).

3.2. Evaluation protocol

The training subset of AVspoof database is used for training a PAD or an ASV system. The development set is used for determining hyper-parameters of the system, including an equal error rate threshold. For that purpose, the system is run on each of the samples from development set, producing scores indicating how similar these samples are to the previously built models (real or spoofed for PAD and client models for ASV). Once such scores for all samples are obtained, knowing the correct correspondence (whether it is spoofed or real data for PAD and a client ID for ASV), we can split the scores in two sets ensuring that false acceptance rate (FAR) and false reject rate (FRR) are equal. This equal rate is usually called equal error rate (EER). The median value of the split scores is the EER threshold, since this is the specific value of the system that leads to EER.

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

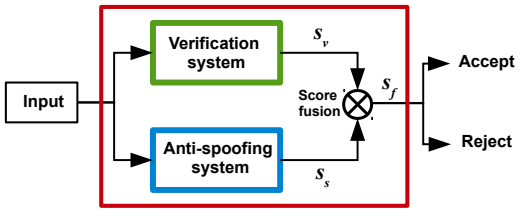


Figure 3: An ASV-PAD joint system based on score fusion.

Applying the EER threshold obtained from *Dev* set to the scores of the *Test* set leads to another pair of FAR and FRR values, which are the measures of the system's performance in uncontrolled evaluation settings. In a perfectly consistent PAD or ASV system, FAR and FRR values on the *Test* set would be the same as FAR and FRR values obtained for *Dev* set. Hence, to summarize the performance of the system in one value, a half total error rate (HTER) is computed as a mean of FAR and FRR. The HTER is then used as an overall measure of the PAD or ASV system's performance.

For more details on the evaluation methodologies for PAD and ASV systems, please refer to [2].

4. Integrating ASV and PAD

As described in Section 2, multiple presentation attack detection systems have been proposed to detect whether a given speech sample is real or spoofed. However, the purpose of a PAD system is to work in tandem with a verification system, so that the joint system can effectively separate the genuine data from both zero-effort impostors (genuine data but incorrect identity) and spoofed attacks (spoofed data for the correct identity).

Typically, a PAD system is sequentially combined with an ASV system, so that spoofed data is filtered out first and only non-spoofed data is passed along for the verification. Such filtering decision can occur at different components of the system (at feature extraction, training, post-processing, etc.), however, in this paper, we combine or fuse these systems at the score level, using a parallel scheme, as is illustrated in Figure 3. Basically, the scores from each system are taken and transformed into one set of scores, which are used to separate real samples that belong to the correct identity from zero-impostor and presentation attacks. This classification task can be performed using different approaches, but we adopt a logistic regression classifier, which results in a straight line separation, as illustrated by the scatter plot in Figure 5a. The classifier is pre-trained on the scores from both ASV and PAD systems, which are computed on the training set of AVspoof dataset.

Considering a score level fusion, it is then important to perform a thorough evaluation of the combined/fused system to understand how incorporating anti-spoofing measures affects verification accuracy for both real and spoofed data. In this paper, we adopt an evaluation methodology specifically designed for performance assessment of fusion system proposed in [2] for face recognition.

Table 1: MFCC and LBP-based PAD systems performance.

Systems	**EER (%)**	**FAR (%)**	**FRR (%)**	**HTER (%)**
MFCC-based	5.66	5.61	8.14	6.88
LBP-based	2.02	1.35	2.96	2.15

To help ASV systems resist spoofing attacks, we consider two presentation attacks detection systems: (i) based on histograms of $LBP_{8,1}$ features with logistic regression classifier, which is an extension of the system presented in [1], and (ii) based on MFCC features with GMM-based classifier, which was evaluated in [11] on 'logical access' attacks of ASVspoof[4] database.

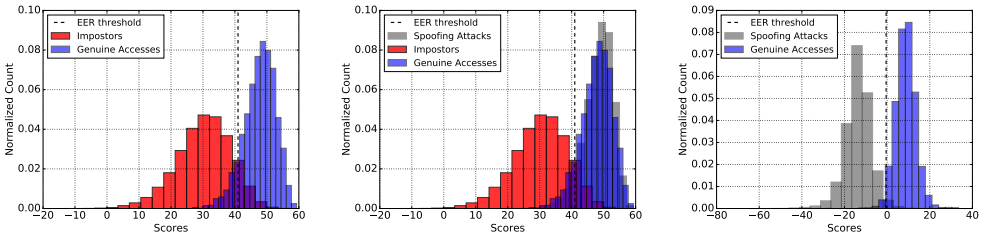
In MFCC-based PAD system, 19 MFCCs with deltas and double-deltas were used as features. Two GMMs are built for real data and spoof attacks and the evaluation score is computed as a log-likelihood between these two GMMs. In LBP-based PAD system, spectrogram (filtered with 40 Mel-scaled filters, log values) is split in two lower and higher bands. For each band, a histogram of regular $LBP_{8,1}$ values is computed and used as the feature. Logistic regression classifier is used to train and evaluate the features.

Table 1 presents the error rates for the considered PAD systems against the presentation attacks of AVspoof database. To have a clearer understanding about the performance of the selected PAD systems, we also plot histograms of score distributions for real data and attacks in both *Dev* and *Test* sets. Figure 4 shows these score distributions for ASV and the best PAD system (LBP-based). The detailed results for MFCC-based system implemented in the BEAT platform[7] can be examined and the detection error tradeoff (DET) curves of both PAD systems can be compared using the online report[8] as well.

5. Experiment results and discussion

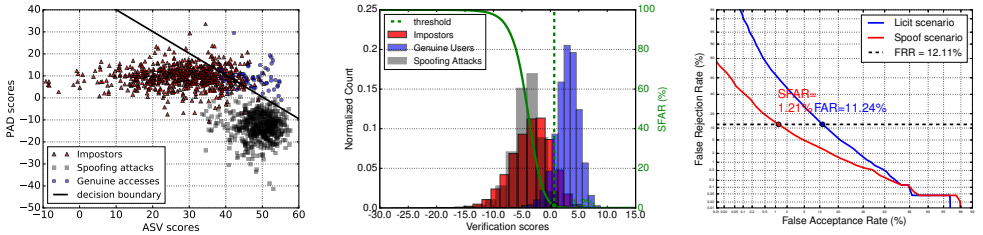
In this section, we demonstrate the vulnerability of stand-alone speaker verification and the improved resistance to spoofing when it is integrated with a PAD system. In addition to two PAD systems presented in Section 4, we consider two ASV systems based on ISV modeling [13] (implemented in BEAT platform extending the system in [3]) and *i-vectors* [5] (stand-alone open source implementation), which are the state of the art speaker verification systems able to effectively deal with intra-class and inter-class variability. In these systems, voice activity detec-

[7] https://www.beat-eu.org/platform/reports/10047079111/
[8] https://www.beat-eu.org/platform/reports/965188923/



(a) ASV system, 'licit' scenario (b) ASV system, 'spoof' scenario (c) PAD system

Figure 4: Histogram distributions for *i-vector* based ASV system in 'licit' and 'spoof' scenario, and LBP-based PAD system.



(a) Scatter plot, Test set (b) Histogram distribution, Test set (c) DET curves, Test set

Figure 5: A scatter plot, histogram distributions, and DET curves for joint *i-vector* and LBP-based system.

tion is based on the modulation of the energy around 4Hz, the features include 19 MFCCs and energy, with their first and second derivatives, and modeling was performed with 256 Gaussian components. In *i-vectors* based system, the dimension of *i-vectors* is 100.

5.1. Vulnerability of ASV to presentation attacks

Table 2 demonstrates how *i-vectors* (first row) and ISV-based (third row) ASV systems perform in two different scenarios: (i) when there are no attacks present (zero-impostors only), referred in the paper as 'licit' scenario, and (ii) when the system is being spoofed with presentation attacks, referred as 'spoof' scenario. Histograms of score distribution in Figure 4b also illustrate the effect of attacks on *i-vectors* based ASV system in 'spoof' scenario, compared to 'licit' scenario in Figure 4a. For more details on ASV-based implementation, please refer to the implementations available online for 'licit'[9] and for 'spoof'[10] scenarios.

From Table 2, it can be noted that both ASV systems perform relatively well under 'licit' scenario with ISV-based system showing lower HTER of 4.19%. However, when a spoofed data is introduced, without a PAD system in place, the spoofing false acceptance rate (SFAR) significantly increases reaching 97.75% for ISV-based and 94.92% for *i-vectors* based systems, which leads to an HTERs of 50.54% and 51.61%, respectively. It means that a typical verification system is not able to correctly distinguish presentation attacks from genuine data.

5.2. Joint operation of ASV and PAD

As presented in Section 4, in score-based fusion of PAD and ASV systems, we make a decision about each speech sample using the scores from both PAD and ASV. The resulted joint system can effectively distinguish genuine data from presentation attacks, as demonstrated in Figure 5b for ASV based on *i-vector* integrated with LBP-based PAD system. Integration of PAD system effectively reduced SFAR from 94.92% that stand-alone *i-vector* ASV had under 'spoof' scenario to SFAR of 1.21%, at the expense of slightly degraded verification performance in 'licit' scenario, when HTER increases from 8.55% to 11.67%, once PAD is added (see Table 2 and Figure 5c).

From the Table 2 and BEAT comparative online report[11] for ISV-based ASV with LBP-based PAD and ISV-based ASV with MFCC-based PAD systems, it is clear that they are less effective in filtering out presentation attacks with HTER values 12.84% and 23.39% respectively. However, it is clear that a joint system that fuses the outcomes of both PAD and ASV systems is more resistant to attacks in a 'spoof' scenario, while still performing well in the 'licit' scenario compared to a stand-alone ASV system.

[9] https://www.beat-eu.org/platform/reports/2044414884/
[10] https://www.beat-eu.org/platform/reports/1686711283/
[11] https://www.beat-eu.org/platform/reports/1229989776/

Table 2: Performance of *i-vector* and ISV-based verification systems individually and jointly with LBP and MFCC-based PAD systems on real and spoofed samples from *Test* set of AVspoof database.

System	Zero-impostors only			Spoofed attacks only		
	FAR (%)	FRR (%)	HTER (%)	SFAR (%)	SFRR (%)	HTER (%)
i-vector based ASV	8.80	8.30	8.55	94.92	8.30	51.61
joint ASV-PAD: *i-vector* and LBP-based	11.24	12.11	11.67	1.21	12.10	6.66
ISV-based ASV	5.23	3.14	4.19	97.75	3.14	50.54
joint ASV-PAD: ISV and LBP-based	8.34	8.16	8.25	12.84	19.52	16.18
joint ASV-PAD: ISV and MFCC-based	5.60	8.31	6.95	23.39	30.48	26.94

6. Conclusion

In this paper, we consider score-based integration of several PAD and ASV systems. Experimental results show a significantly increased resistance of joined ASV-PAD systems to presentation attacks of AVspoof database (e.g., false acceptance rate for the attacks drops from above 90% for stand-alone ASV to less than 2% for an *i-vector* and LBP-histograms based ASV-PAD joint system) at the expense of slightly degraded performance for the scenario with no attacks. An important contribution is also an open source and online-based implementations of the considered ASV, PAD, and joint ASV-PAD systems.

In the future work, we will consider a cascading scheme for score fusion and compare it with the presented parallel scheme. We will also focus on the development of novel presentation attack databases and on exploring multi-modal systems, when both ASV and PAD systems of different modalities, e.g., speech and image, are integrated to improve the performance in both 'licit' and 'spoof' scenarios.

Acknowledgements

This work was conducted in the framework of EU funded BEAT project (#284989) and was partially funded by Swiss Centre for Biometrics Research and Testing. The authors would also like to thank Tiago de Freitas Pereira and André Anjos for the help with BEAT platform implementation.

References

[1] F. Alegre, A. Amehraye, and N. Evans. A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–8, Arlington, VA, Sept. 2013. 1, 2, 4

[2] I. Chingovska, A. Anjos, and S. Marcel. Biometrics evaluation under spoofing attacks. *IEEE Trans. on Information Forensics and Security*, 9(12):2264–2276, Dec 2014. 1, 4

[3] T. de Freitas Pereira and S. Marcel. Periocular biometrics in mobile environment. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–7, Arlington, VA, Sept. 2015. 4

[4] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga. Evaluation of speaker verification security and detection of HMM-based synthetic speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8):2280–2290, Oct. 2012. 2

[5] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, May 2011. 1, 3, 4

[6] ISO/IEC JTC 1/SC 37 Biometrics. DIS 30107-1, information technology – biometrics presentation attack detection. American National Standards Institute, Jan. 2016. 2

[7] A. Janicki. Spoofing countermeasure based on analysis of linear prediction error. In *Interspeech*, pages 2077–2081, Dresden, Germany, Sept. 2015. 2

[8] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1435–1447, May 2007. 3

[9] S. Kucur Ergunay, E. Khoury, A. Lazaridis, and S. Marcel. On the vulnerability of speaker verification to realistic voice spoofing. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–6, Arlington, VA, Sept. 2015. 1, 2, 3

[10] T. B. Patel and H. A. Patil. Combining evidences from Mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. In *Interspeech*, pages 2062–2066, Dresden, Germany, Sept. 2015. 2

[11] M. Sahidullah, T. Kinnunen, and C. Hanilçi. A comparison of features for synthetic speech detection. In *Interspeech*, pages 2087–2091, Dresden, Germany, Sept. 2015. 2, 4

[12] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui. Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification. In *Interspeech*, pages 239–243, Dresden, Germany, Sept. 2015. 2

[13] R. Vogt and S. Sridharan. Explicit modelling of session variability for speaker verification. *Comput. Speech Lang.*, 22(1):17–38, Jan. 2008. 1, 3, 4

[14] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li. Spoofing and countermeasures for speaker verification: A survey. *Speech Communication*, 66:130–153, Feb. 2015. 1, 2

[15] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Interspeech*, pages 2037–2041, Dresden, Germany, Sept. 2015. 2