Spoofing Detection for Personal Voice Assistants

Arun Sankar M. S.
School of Computer Science and Information Technology (CSIT), University College Cork
Cork, Ireland
asankar@ucc.ie

Phillip L. De Leon
Department of Electrical Engineering, University of Colorado Denver
Denver, U.S.A.
Phillip.DeLeon@ucdenver.edu

Utz Roedig
School of Computer Science and Information Technology (CSIT), University College Cork
Cork, Ireland
u.roedig@ucc.ie

ABSTRACT

Personal Voice Assistants (PVAs) are common acoustic sensing systems that are used as a speech-based controller for critical systems making them vulnerable to speech spoofing attacks. Prior research has focused on the discrimination of genuine and spoofed speech for applications with large population speaker verification and challenges such as ASVspoof have advanced this work over the last few years. In this paper, we consider spoofing detection in a PVA setting where the number of household users is small. We show that when pre-trained models are adapted to household users, spoofing detection is improved. Furthermore, we demonstrate that adaptation is still effective in realistic scenarios where only genuine speech of household users is available but the generation of spoofed speech samples for household users is undesirable.

CCS CONCEPTS

• **Security and privacy → Systems security**.

KEYWORDS

Computer security, Acoustic sensing, Biometrics, Speaker recognition, Speech processing

1 INTRODUCTION

Using voice in Internet of Things (IoT) devices has revolutionized the way we interact with technology and our surroundings. Personal Voice Assistants (PVAs) are increasingly used as interfaces for digital environments such as smartphones, home appliances and vehicles [3]. PVAs are both IoT devices themselves, performing acoustic sensing, and central hubs for controlling other IoT devices, creating a more interconnected and convenient smart home experience. A PVA can be a compact, standalone device built for a specific application such as a smart speaker (e.g. Amazon Echo or Google Home) or it can be integrated within a device that has speech processing capabilities such as a smartphone or car navigation system (e.g. Siri). In most cases, the PVA is split into a front-end located on the device, consisting of a microphone, speaker, and limited computing capabilities to detect a wake word and a back-end located remotely in the cloud to process user requests. As we increasingly rely on PVAs it is necessary to consider Automatic Speaker Verification (ASV) in this context to provide authentication or improved user experience. Current commercially available PVAs are starting to integrate ASV within products, e.g. Alexa Voice ID.

ASV is a low-cost, convenient, and accurate technology for biometric authentication and has been the subject of research for several decades [12]. ASV systems are known to be vulnerable to spoofing attacks via impersonation, replay, speech synthesis, twins, and voice conversion [5, 6, 22]. Countermeasures to detect spoofed speech and thus prevent an attack, have been proposed and remain in active development [20]. These countermeasures classify genuine versus spoofed speech based on features such as Constant-Q Cepstrum Coefficients (CQCC), Linear-Frequency Cepstrum Coefficients (LFCC), neural network embedding, and statistical features of Instantaneous Amplitude (IA) and Instantaneous Frequency (IF) [1, 10, 13, 25]. Organized trials and evaluations such as the ASVspoof challenge initiated in 2015 and most recently in 2019 and 2021, have assisted with advancing the research [2, 21].

Within a PVA, the vulnerability of ASV to spoofing attacks poses a serious security threat. In this paper, we examine how best to integrate a spoofing detector in a PVA environment. The various types of spoofing detectors proposed so far have been developed without considering specific applications for which they have to be used. For example in a PVA application, the speaker verification or identification is limited to a very small number of enrolled users which is often much smaller than other applications such as ASV for banking applications. Thus any spoofing detector developed for a PVA need only safeguard the system from PVA users' spoofed speech and not other speakers in general. In such a scenario, the spoofing detector could be highly tuned to the small group of PVA users, which may increase detection accuracy due to the reduction of inter-speaker variability. This work investigates the following. (i) Which is a better approach in a PVA setting—a speaker-independent spoofing detector or a spoofing detector adapted to only PVA users? (ii) If adaptation is beneficial, what are the different adaptation approaches with respect to PVA and non-PVA users' genuine and spoofed speech and are these approaches practical?

The contributions of this work are:

(1) We demonstrate that adaptation of a speaker-independent, pre-trained spoofing detector to PVA users improves detection performance.
(2) We provide an analysis of the adaptation spectrum in terms of data sources, ranging from ideal to more practical scenarios.
(3) We show that adaptation can be performed using only genuine speech of PVA users in combination with spoofed speech

Sensors S&P '23, November 12–17, 2023, Istanbul, Turkiye

ACM ISBN 979-8-4007-0440-6/23/11…$15.00
https://doi.org/10.1145/3628356.3630114

of non-PVA users, representing a practical scenario that does not require spoofed speech generation of PVA users.

The remaining paper is structured as follows. Section 2 describes the motivation for doing this work and the details of ASVspoof challenge and database are given in section 3. The experimental layout is detailed in section 4 and the results are discussed given in section 5. Section 6 concludes the paper.

2 BACKGROUND AND MOTIVATION

PVAs use speaker recognition either to verify a claim of identity based on a voice sample or to identify the speaker from the set of authorized users, typically those which reside in the household [3]. Apple's Siri uses Speaker Verification (SV) to authenticate the user after which the PVA gets triggered and provides access to the user. Other PVAs such as Google Assistant may be triggered via a wake word by an unknown speaker, i.e. a person from outside the household. Both Amazon and Google support Speaker Identification (SI) to provide personalized services that match the profile of the speaker. Unauthorized access to the services of the PVA may pose a threat and can be dealt with using Speaker Recognition (SR) and spoofing detection [3]. The latter has been of interest to researchers for some time now as a wide range of voice spoofing systems exist, e.g. replay or Text-to-Speech (TTS) which may "pass" the SR system [18].

The various logical access (LA) spoofing detection methods developed differ by the front-end features used to acquire discriminative information and by the back-end classifiers used for generating the decision score based on which genuine/spoof speech classification is performed. The promising features used for spoofing detection are especially but not limited to CQCC, LFCC, Mel-Frequency Cepstrum Coefficient (MFCC), Inverse Mel-Frequency Cepstrum Coefficients (IMFCC), and neural network embedding [4, 10, 13, 17, 25]. In some mechanisms, the above-mentioned features are used in combination with source features such as epochs, peak to side lobe ratio to obtain the complementary information that aids detection [7, 16, 23]. The conventional feature extraction is carried out in the frequency domain using filter banks to obtain the short-term sub band spectral features. Some Deep Neural Networks (DNN) based spoofing detection methods use these features to extract the network embeddings that serve as the feature for categorization [15].

To the best of our understanding, most research in spoofing detection can be traced to earlier work [6] and through more recent trials including ASVspoof 2019 and ASVspoof 2021, considers the development of speaker-independent detectors for relatively large-scale applications such as SV. The PVA presents a subtly different problem in that the spoofing detector need not be speaker-independent since only a small set of speakers are authorized to use the PVA. When the spoofing detector is used in PVA applications, we accept the classification of genuine speech as spoofed, provided this is not the genuine speech of the PVA users. Although acceptable in the PVA case, this is not an acceptable result from the classic spoofing detector problem. The small set of PVA users in this application may improve overall (spoofing detection and SR) accuracy in three ways: 1) as is well known when the number of speakers in a SR system is small, accuracy increases; 2) since the spoofing detector need not be speaker-independent, detection accuracy may also be increased due to a reduction in the within-class variability; and 3) rejection of genuine speech of non-PVA users by the spoofing detector is acceptable. These differences should in theory result in much better spoofing detection rates due to the smaller subspace of users in the PVA application.

3 ASVSPOOF CHALLENGE DATA SET AND EVALUATION METRIC

The ASVspoof challenge series was initiated in 2015 with the motivation of advancing spoofing detection and countermeasures [21]. The first challenge was focused on voice conversion and synthetic speech attacks while the second spoof challenge organized in 2017 concentrated on replay attacks as they are much easier to generate without any technical expertise. The third spoof challenge took place in 2019 and considered speech synthesis, voice conversion, and replay attacks. The fourth challenge organized recently in 2021 focused on discriminating between genuine and spoofed or deepfake speech using ASVspoof 2019 database [2].

The ASVspoof 2019 challenge database consists of a LA partition containing voice conversion and speech synthesis examples in addition to the physical access (PA) partition which contains replay examples. Each partition contains training, development and evaluation subsets. The training and development subsets are used for conducting experiments related to the development of the detection model while the evaluation set is utilized for measuring the detection performance of the developed model. The training and development subsets of LA contain 6 spoofing attacks which are considered as known attacks and used for the construction of the detection model. The evaluation subset of LA has 2 known attacks and 11 unknown attacks to determine the efficiency of the developed model on attacks that are unknown to the system or in other words on attacks that are not used for training the model. In addition, each subset also contains examples of human-produced speech. All speech examples, including the source utterances for creating the spoofed speech, are taken from the VCTK corpus [19]. Details of the database are summarized in Table 1.

The training and development data sets are built using the same set of spoofing attacks (A01-A06). Spoofing attacks A01 to A04 are based on TTS methods while attacks A05 and A06 use voice conversion (VC) methods. The attacks A01-A03 are neural network based TTS systems and attack A04 does TTS using the waveform concatenation method. The evaluation data set consists of 13 spoofing attacks (A07-A19) out of which only 2 attacks (A16 and A19) are known attacks. Attacks A16 and A19 use the same spoofing techniques as attacks A04 and A06 respectively. The unknown attacks consist of six TTS based methods (A07-A12), two VC methods (A17 and A18) and three hybrid models (A13-A15). The hybrid models use a combination of VC and TTS for the generation of spoofed speech.

The following metrics are used for quantifying the detection performance of the spoofing detector.

- *Equal Error Rate (EER)* - An ideal spoofing detector should flag spoofed speech and pass genuine speech but in reality there is always some error which is quantified using False Acceptance Rate (FAR) and False Rejection Rate (FRR).

Table 1: Description of the logical access partition of the ASVspoof 2019 challenge database. In the training and development sets, all six spoofing attacks are present in the spoofed examples. In the evaluation set, A16 is the same as A04 and A19 is the same as A06.

Database attributes	Training set	Development set	Evaluation set
Spoofing attack algorithms	A01-A06	A01-A06	A07-A19
Spoofing methods	TTS (4) VC (2)	TTS (4) VC (2)	TTS (6) VC (2) Hybrid (3)
Known attacks	6	6	2 (A16, A19)
Unknown attacks	0	0	11
No. of genuine examples	2580	2548	7355
No. of spoofed examples	22800 (3800×6)	22296 (3716×6)	63882 (4914×13)
No. of male speakers	8	(4+4)	21
No. of female speakers	12	(6+6)	27

The False Acceptance Rate is the ratio of spoofed speech samples wrongly classified as genuine speech and False Rejection Rate is the ratio of genuine samples misclassified as spoofed speech.

It is desirable to minimize both FAR and FRR for improving the efficiency of detection systems. But adjusting the detection threshold to reduce either of the errors harm the other. The detection threshold plot has a point where both the error rates are equal and that common value is called the EER which is considered a metric in ASV spoof 2019 challenge.

- *tandem-Detection Cost Function:* The EER metric is sufficient to quantify the performance of a stand alone spoofing detector. But when this detector is integrated into an ASV system, the impact of countermeasure on verification performance cannot be evaluated by EER metric. In such a scenario, the tandem-Detection Cost Function (t-DCF) metric [8] measures the impact of spoofing and countermeasure on the reliability of ASV system by combining the verification and spoofing errors. The minimum normalized tandem-Detection Cost Function is expressed in the form

$$\text{t-DCF}_{\text{min}} = \min_{Thr} \{\beta P_{cmMISS}(Thr) + P_{cmFAR}(Thr)\}\,. \quad (1)$$

The parameter β depends on the spoofing prior and cost parameters and on miss and false alarm rates of speaker verification. $P_{cmMISS}(Thr)$ and $P_{cmFAR}(Thr)$ are the false alarm and miss rates of the countermeasure at threshold Thr.

For additional information, please see [2].

4 EXPERIMENTS AND EVALUATION

In this section, we describe the experiments conducted to analyze the adaptation of a generalized spoofing detector to PVA users. Experiments are conducted in two ways: (i) adaptation of a pre-trained model only to the speech of PVA users with known spoofing attacks and (ii) adaptation of a pre-trained model to the speech of PVA users and to unknown attacks. The latter experiment allows investigation of system performance when presented with spoofed speech using a technique unknown to the system while the former allows investigation of system performance with knowledge of all techniques used to spoof speech. In a practical system, the former takes into consideration the practical and realistic scenario where generating the spoofed speech of PVA users is difficult and would require updates to the model as new spoofing attacks become known.

For this work we use the ASVspoof 2019 data set which consists of training and development sets for building spoofing detector models and an evaluation set to measure performance. We note that the ASVspoof 2021 challenge has the same training and development sets as the 2019 challenge but the evaluation set differs [11].

For adapting the spoofing detector to PVA users, we begin with a pre-trained model based on Rawnet2 which is among the top performing detectors among four baseline classification systems of the ASVspoof 2021 challenge [24]. Rawnet2 is a DNN framework that uses raw unprocessed speech signals as input. This is processed by a set of Mel-scale sinc filters followed by residual blocks and Gated Recurrent Unit (GRU) for feature extraction. The feature embeddings are given to a fully connected layer with a softmax activation function to yield the binary classification of genuine or spoofed speech. For complete details on Rawnet2, we refer the reader to [14]. Spoofing detection performance is measured in the ASVspoof 2019 challenge using EER, which provides the standalone performance of the spoofing detector and t-DCF which measures the spoofing detection performance in combination with the performance of the ASV system. While developing the pre-trained model or adapting the pre-trained model to a set of PVA users, the experiments on Rawnet2 are conducted with learning rate, number of epochs, and batch size respectively 0.001, 20, and 32. For the ASVspoof 2019 challenge, Rawnet2 has an EER of 5.64% and t-DCF of 0.1301% on the evaluation set [14].

4.1 ASVspoof 2019 Data Usage

ASVspoof 2019 provides *training*, *development*, and *evaluation* data sets containing genuine speech samples and spoofed speech samples created with different spoofing attacks A01 - A19. The training set consists of both genuine and spoofed speech samples for spoofing attack set $S_0 = \{A01, \ldots, A06\}$ consisting of 20 speakers from LA_0079 to LA_0098. The development set uses the same attack set S_0 but 20 different speakers from LA_0069 to LA_0078 and LA_0099 to LA_0108. The ASVspoof evaluation set consists of genuine and spoofed speech samples for attack set $S_3 = \{A07, \ldots, A19\}$ from speakers LA_0001 to LA_0048. We define the spoofing sets S_1 and S_2 which are subsets of S_3 for conducting experiments. The spoofing set $S_1 = \{A16, A19\}$ which are the same as A04 and A06 present in the training and development sets. Set $S_2 = \{A07, A10, A12, A13, A14, A17\}$ consists of six attacks, three are based on TTS (A07, A10 and A12) and three are combined TTS and VC system (A13, A14 and A17).

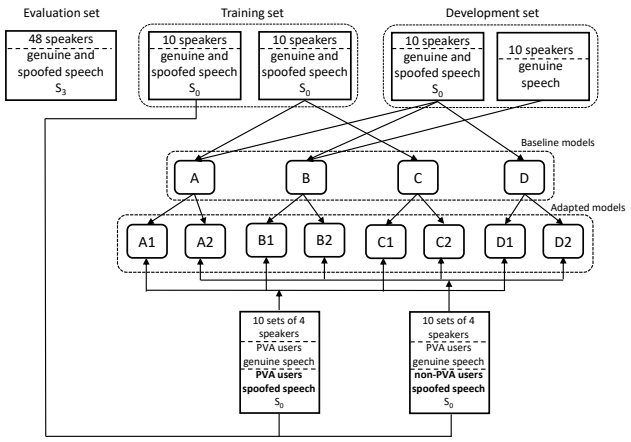

Figure 1: The usage of ASVspoof data to develop baseline models and to further adapt them to PVA users. The spoofing attack sets $S_0 = \{A01, \ldots, A06\}$ and $S_3 = \{A07, \ldots, A19\}$.

Finally, we train/adapt and evaluate the models using ten sets of four randomly chosen PVA users such that we span (F)emale and (M)ale combinations: FFFF, FFFM (2), FFMM (4), FMMM (2), and MMMM. The choice of four users reflects thea rough average of the number of people in a household or household size [9].

4.2 Adaptation to PVA users

Ten speakers from the training data set are selected as PVA users. The speech samples of the remaining 10 speakers of the training data set and the complete development data set are used in various combinations for building pre-trained (baseline) models A - D. The need for these various baseline models is to determine and quantify the extent of model adaptation using speech of PVA users. Figure 1 illustrates how the ASVspoof 2019 data is partitioned and used in developing the pre-trained and adapted models. The pre-trained models are adapted to genuine and/or spoofed speech of PVA users with only known spoofing attacks. The adaptation is done in two ways as shown in Figure 1: (i) using both genuine and spoofed speech of PVA users for models A_1 - D_1 and (ii) using genuine speech of PVA users and spoofed speech of non-PVA users for models A_2 - D_2. The second method of adaptation is chosen by taking into consideration the practical and realistic scenario where generating the spoofed speech of PVA users is difficult.

These pre-trained models are described below.

Model A: Uses genuine and spoofed speech samples of 20 speakers, ten from training and ten from development data sets.

Model B: Uses the entire development data set which has genuine and spoofed speech samples of 20 speakers.

Model C: Uses genuine and spoofed speech samples of 10 speakers selected from the training data set.

Model D: Uses genuine and spoofed speech samples of 10 speakers selected from the development data set.

The training data used for developing models *A* and *B* is nearly twice the size of data used for developing models *C* and *D*. The ten sets of PVA users of group size four are randomly selected from 10 speakers of the training data set which are not used in the development of pre-trained models. The pre-trained/adapted models are evaluated using these sets of PVA users and their average values are shown in Table 2. For each set of PVA users, the non-PVA users are selected from the remaining speakers of PVA users who are not belonging to that particular set.

Table 2: Spoofing detection performance for models adapted from baseline models (A - D). The adaptation is done using PVA users' genuine speech and spoofed speech of PVA users (models A_1 - D_1) or non-PVA users (models A_2 - D_2). The spoofed speech samples for training and adaptation are created using set S_0.

Spoofing Detector Model	PVA users genuine speech	PVA users spoofed speech	non-PVA users spoofed speech	EER (%)
A				0.322
A_1	✓	S_0		0.000
A_2	✓		S_0	0.334
B				3.638
B_1	✓	S_0		0.049
B_2	✓		S_0	0.421
C				1.213
C_1	✓	S_0		0.124
C_2	✓		S_0	0.718
D				6.077
D_1	✓	S_0		0.297
D_2	✓		S_0	1.707

4.3 Adaptation to PVA users and unknown attacks

Training and development sets are used to generate the pre-trained model, i.e. model E which is then adapted using data from the evaluation set creating models F - J. Figure 2 illustrates how the ASVspoof 2019 data is partitioned and used in developing the pre-trained and adapted models with unknown attacks. The model specifics are explained in more detail below. Models F - H are the baseline model adapted with the PVA users' genuine and spoofed speech using various spoofing attacks and models I and J are the baseline model adapted with the PVA users' genuine speech and additional examples of spoofed speech from non-PVA users as shown in Figure 2. None of the speakers in the training and development set are considered as PVA users in these experiments; PVA user groups are built from samples in the evaluation set which are then used for model adaptation. The non-PVA users, required for models I and J, are selected from the remaining speakers in the evaluation set after creating the group of PVA users.

Model E: This is the baseline model developed using Rawnet2, generated using ASVspoof 2019 training and development sets respectively for training and validation purposes. All other models described below are adapted from this baseline model. This model represents what might be the default system on the PVA before any user adaptation.

Model F: Model E is adapted with PVA users' genuine speech and spoofed speech for set S_1. In this model, it is assumed that the

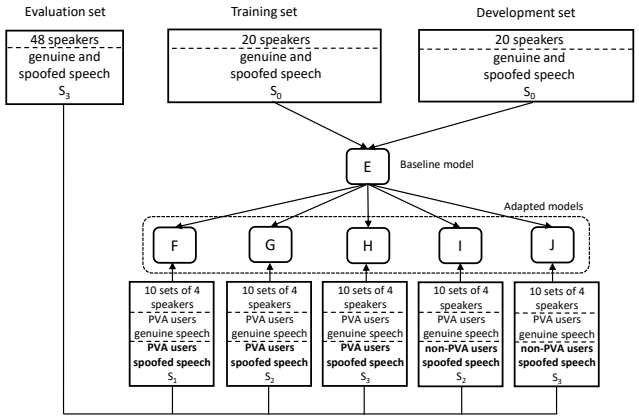

Figure 2: The usage of ASVspoof data to develop baseline models and to further adapt them to both PVA users and unknown attacks. The spoofing attack sets $S_1 = \{A16, A19\}$, $S_2 = \{A07, A10, A12, A13, A14, A17\}$ and $S_3 = \{A07, \ldots, A19\}$.

spoofed speech of PVA users are available only for attacks known at the time the baseline was constructed (known attacks). We note that only PVA users' spoofed speech from S_1 is used which is a subset of S_0 used for training the baseline model. The adaptation only considers household users and not new attacks (attacks not considered in baseline training).

Model G: Model E is adapted with PVA users' genuine speech and spoofed speech for set S_2. It is assumed that some previously unknown attacks are now known and that spoofed speech of PVA users for these attacks is available. This model is adapted considering household users as well as additional attacks.

Model H: Model E is adapted with PVA users' genuine speech and spoofed speech for set S_3. The assumption is that the PVA users' spoofed speech for all the attacks of the evaluation set (S_3) are available for adaptation. It may be considered the "ideal" case in the sense of having a spoofing detector which is highly tuned to the PVA users and spoofing attacks. This model is similar to model G but considers additional spoofing attacks not used in model G.

Model I: Model E is adapted with PVA users' genuine speech and non-PVA users' spoofed speech for set S_2. The non-PVA users are chosen from the speakers of the evaluation set after selecting the PVA users. This model is the same as model G with the difference that the spoofed speech is from non-PVA users.

Model J: Model E is adapted with PVA users' genuine speech and non-PVA users' spoofed speech for set S_3. This model is similar to the model I by using non-PVA users' spoofed speech but adapted to all spoofing attacks of the evaluation set S_3.

The models F - H require systems for generating spoofed speech of the PVA users which requires additional computation and may raise privacy concerns. The models I and J are explored as a practical result of producing a more accurate spoofing detector and they may be considered the most realistic and practical cases.

For models F - H, we partition the evaluation set using 80% of the examples for training and 20% for evaluation; for models I and J, we similarly partition the evaluation set in a similar manner except that we use spoofed speech from four of the remaining speakers of the evaluation set after choosing the PVA users. The baseline and adapted models are evaluated using the test set and the results are shown in Table 3. The results given in the next section are the average EER obtained for the ten sets which are named the PVA users' test set.

Table 3: Spoofing detection performance for models adapted from a baseline model. The baseline model (E) is trained using non-PVA users' genuine speech and attack set S_0 from ASVspoof 2019 training set and represents an "out-of-the-box" system. Models F - J adapt model E with PVA users' genuine speech and spoofed speech for PVA users or non-PVA users from different attack sets, i.e. S_1, S_2, or S_3 from the ASVspoof 2019 evaluation set.

Spoofing Detector Model	PVA users genuine speech	PVA users spoofed speech	non-PVA users spoofed speech	EER (%)
E				4.811
F	✓	S_1		4.228
G	✓	S_2		2.694
H	✓	S_3		0.928
I	✓		S_2	3.901
J	✓		S_3	1.706

5 RESULTS AND DISCUSSION

In this section we describe the spoofing detection performance of the various models given in the previous section on PVA users' test set for both experiments.

Average EER values for baseline models adapted to PVA users' genuine and/or spoofed speech on PVA test data set are shown in Table 2. The t-DCF metric is normally used to evaluate the impact of spoofing and its countermeasure on ASV system but is not appropriate in this work because of the differences given in Section 2 for spoofing detection in the PVA problem. The accuracy of baseline models A - D varies due to the amount and type of training data used in the development of the models. The difference in the amount of training data is the reason why baseline models A and B have lower EER than models C and D respectively. Even though models A and B use nearly the same number of speakers and amount of data for training, EER of A is much lower than B due to the inclusion of speech samples from the training data set for building the model. This is because Rawnet2 is developed for spoofing detection with the objective of giving the best performance on the development set while built using the training data set. This explains why model C performs better than model D even though both models have nearly the same amount of speaker data for training.

From the results shown in Table 2, the adaptation improves the spoofing detection performance of all baseline models with the exception of model A. The maximum performance is obtained while adapting the baseline model to both genuine and spoofed speech of PVA users (models A_1 - D_1). However, the adaptation using PVA users' genuine speech and non-PVA users' spoofed speech (models A_2 - D_2) gives better detection performance than its baseline model.

The EER values obtained for various models on PVA test data set while adapting to PVA users' speech and new spoofing attacks are summarized in Table 3. *Model E:* This baseline model is trained with non-PVA users' genuine speech and non-PVA users' spoofed speech from attack set S_0 ={A0, …, A6} of ASVspoof 2019 training set as described in the previous section. This model has an EER of 4.811% which is different from the 5.64% EER reported in [14] for the Rawnet2 system on the evaluation set of ASVspoof 2019 challenge. This difference is due to our test set consisting of only four speakers, i.e. PVA users as described in the previous section.

Model F: Results show that adaptation of model E to the genuine and spoofed speech of PVA users slightly improves EER from 4.811% to 4.228%. The EER improvement is only attributed to the fact that the relevant speakers are considered. It also has to be noted that the number of samples and known spoofing attacks available for adaptation was small (160 samples per speaker per spoofing attack); a larger data set would likely improve the EER.

Model G: This model is obtained by adapting model E not only to the genuine and spoofed speech of PVA users but also to a new set of attacks S_2. This combined effect reduced the EER from 4.811% to 2.694%. We note that by adapting model E to nearly half of the unknown attacks in the test set, the spoofing detection performance is significantly improved. In comparison with model F, the additional attacks used in generating model G, i.e. S_2 further contributes to the reduction of EER from 4.228% to 2.694%.

Model H: While adapting the model E to the genuine and spoofed speech of PVA users and to the full set of spoofing attacks in the test set S_3, the spoofing detection performance is drastically improved by reducing the EER from 4.811% to 0.928%. This is the ideal situation in adaptation where there is a perfect match between the adaptation set and test set for spoofing attacks and speakers. Furthermore, the exposure of model E to all unseen attacks of the test set which is additionally done in this model compared to model F reduced the EER from 4.228% to 0.928% which is the lowest EER obtained with adaptation in this work. Thus the efficiency of spoofing detectors can be improved by adapting to new types of spoofing attacks. Here the spoofing detection performance improvement is not only due to the adaptation of the baseline model to the complete set of spoofing attacks but also due to the familiarization with both genuine and spoofed speech of PVA users.

Model I: So far the adaptation of model E is done using both genuine and spoofed speech of PVA users but here non-PVA users' spoofed speech is used along with PVA users' genuine speech for adaptation. By doing so, the EER is reduced from 4.811% to 3.901% where the system is exposed to nearly half of the unknown attacks in the test set S_2. If the spoofed speech of PVA users are used in place of non-PVA users, the EER could have reduced from 3.901% to 2.694% which is same as model G. This shows the importance of adapting the baseline model to the genuine and spoofed speech of PVA users. Even if the spoofed speech of PVA users is not available or impractical to obtain, using non-PVA users' spoofed speech improves detection performance beyond the baseline model and is also better than the model F, which is adapted to the spoofed speech of PVA users for known attacks.

Model J: In this case, by using examples of additional attacks in the adaptation, the EER is further reduced from 3.901% (model I) to 1.706%. Compared to model E, model J includes adaptation to the PVA users' genuine speech and the complete set of attacks S_3 and results in EER reducing from 4.811% to 1.706% which is a reduction nearly by a factor of 3. In model H we use the complete set of attacks S_3 of PVA users whereas Model J uses the complete set of attacks S_3 of non-PVA users which illustrates the trade-off in using PVA users' spoofed speech.

All models F to J adapted to PVA users show an improvement in terms of EER compared to the baseline model (model E). Improvement is observed independent of the specific data considered for adaptation; improvements are present if PVA users' genuine and spoofed speech or only PVA users' genuine speech is considered. Similarly, improvements are observed regardless if data considering new attack forms or not is used. Thus, we conclude that model adaptation to PVA users is always beneficial. In a practical PVA setting, the baseline model E can be considered as the default "out-of-the-box" spoofing detector shipped with the PVA. This model should then be adapted to the PVA users to improve performance.

The best adaptation can be achieved using PVA users' genuine speech and spoofed speech (see models F to H). However, to use this approach in practice requires the generation of spoofed speech of PVA users. This may not be possible for two reasons. First, generation of spoofed speech requires computational resources which may not be available. Second, users may object to the generation of spoofed speech; they may feel uncomfortable with a system producing fake samples of their speech.

Our results show that it is possible to use PVA users' genuine speech and non-PVA users' spoofed speech for model adaptation. This is practical as it is easy to collect speech samples from household PVA users while for spoofed speech a pool of centrally generated spoofed speech is used. This removes the need to generate spoofed samples of users addressing their security concerns.

6 CONCLUSIONS

In this paper, we considered spoofing detection in a PVA setting where the number of household users is small. Our experiments show that adaptation of a pre-trained model to the PVA users' genuine speech improved spoofing detection. We also considered adaptation of the pre-trained model to both, PVA and non-PVA users' spoofed speech with various sets of spoofing attacks. In the ideal case, adapting the pre-trained model using PVA users' genuine speech and spoofed speech results in the lowest EER. However, due to necessary computational resources and privacy issues, adapting with PVA users' spoofed speech may not be practical. This situation can be overcome by adapting the model using PVA users' genuine speech but non-PVA users' spoofed speech, resulting in an EER as low as 1.706% which is significantly lower than the baseline results. Future work will focus on the combined performance of this spoofing detector with ASV system in a PVA environment.

7 ACKNOWLEDGEMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 19/FFP/6775 and 13/RC/2077_P2. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

REFERENCES

[1] M. S. Arun Sankar, Phillip L. De Leon, Steven Sandoval, and Utz Roedig. 2022. Low-Complexity Speech Spoofing Detection using Instantaneous Spectral Features. In *Proc. Int. Conf. Systems, Signals and Image Process. (IWSSIP)*, Vol. CFP2255E-ART. 1–4.

[2] ASVspoof Consortium. 2019. ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan. https://www.asvspoof.org/asvspoof2019/asvspoof2019_evaluation_plan.pdf. [Online; accessed 21-September-2023].

[3] Peng Cheng and Utz Roedig. 2022. Personal Voice Assistant Security and Privacy—A Survey. *Proc. IEEE* 110, 4 (April 2022), 476–507.

[4] Bhusan Chetttri et al. 2019. Ensemble Models for Spoofing Detection in Automatic Speaker Verification. In *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. 1018–1022.

[5] Phillip L. De Leon, Vijendra Raj Apsingekar, Michael Pucher, and Junichi Yamagishi. 2010. Revisiting the security of speaker verification systems against imposture using synthetic speech. In *Proc. IEEE Int. Conf. Acoust. Speech Sig. Process. (ICASSP)*. 1798–1801.

[6] Phillip L De Leon, Michael Pucher, Junichi Yamagishi, Inma Hernaez, and Ibon Saratxaga. 2012. Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech. *IEEE Trans. Audio, Speech, Language Process.* 20, 8 (Oct. 2012), 2280–2290.

[7] Sarfaraz Jelil, Rohan Kumar Das, S. Prasanna, and R. Sinha. 2017. Spoof Detection Using Source, Instantaneous Frequency and Cepstral Features. In *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. 22–26.

[8] Tomi Kinnunen, Kong Aik Lee, Hector Delgado, Nicholas Evans, Massimiliano Todisco, Md Sahidullah, Junichi Yamagishi, and Douglas A. Reynolds. 2019. t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification. arXiv:1804.09618 [eess.AS]

[9] Stephanie Kramer. 2020. With billions confined to their homes worldwide, which living arrangements are most common? https://pewrsr.ch/2w36OXH. [Online; accessed 19-October-2023].

[10] Galina Lavrentyeva et al. 2019. STC Antispoofing Systems for the ASVspoof2019 Challenge. In *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. 1033–1037.

[11] Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi H. Kinnunen, Ville Vestman, Massimiliano Todisco, Héctor Delgado, Md Sahidullah, Junichi Yamagishi, and Kong Aik Lee. 2019. ASVspoof 2019: The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database. https://doi.org/10.7488/ds/2555. [Online; accessed 19-July-2023].

[12] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. 2000. Speaker verification using adapted Gaussian mixture models. *Digital signal processing* 10, 1-3 (Jan. 2000), 19–41.

[13] Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Hong Yu, Tomi Kinnunen, Nicholas Evans, and Zheng-Hua Tan. 2016. Integrated Spoofing Countermeasures and Automatic Speaker Verification: An Evaluation on ASVspoof 2015. In *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. 1700–1704.

[14] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. 2021. End-to-End anti-spoofing with RawNet2. In *Proc. IEEE Int. Conf. Acoust. Speech Sig. Process. (ICASSP)*. 6369–6373.

[15] Hemlata Tak, Jee weon Jung, Jose Patino, Massimiliano Todisco, and Nicholas W. D. Evans. 2021. Graph Attention Networks for Anti-Spoofing. In *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. 2356–2360.

[16] Xiaohai Tian, Xiong Xiao, Eng Siong Chng, and Haizhou Li. 2016. Spoofing Speech Detection Using Temporal Convolutional Neural Network. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. 1–6.

[17] Massimiliano Todisco, Héctor Delgado, Kong Aik Lee, Md Sahidullah, Nicholas Evans, Tomi Kinnunen, and Junichi Yamagishi. 2018. Integrated Presentation Attack Detection and Automatic Speaker Verification: Common Features and Gaussian Back-end Fusion. In *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. 77–81.

[18] Massimiliano Todisco et al. 2019. ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection. In *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. 1008–1012.

[19] C. Veaux, J. Yamagishi, and Kirsten MacDonald. 2017. VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit. https://datashare.ed.ac.uk/handle/10283/3443. [Online; accessed 19-October-2023].

[20] Zhizheng Wu, Phillip L. De Leon, Cenk Demiroglu, Ali Khodabakhsh, Simon King, Zhen-Hua Ling, Daisuke Saito, Bryan Stewart, Tomoki Toda, Mirjam Wester, and Junichi Yamagishi. 2016. Anti-Spoofing for Text-Independent Speaker Verification: An Initial Database, Comparison of Countermeasures, and Human Performance. *IEEE Trans. Audio, Speech, Language Process.* 24, 4 (April 2016), 768–783.

[21] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov. 2015. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. 2037–2041.

[22] Zhizheng Wu and Haizhou Li. 2015. On the Study of Replay and Voice Conversion Attacks to Text-Dependent Speaker Verification. *Multimed. Tools Appl.* 75, 3 (Dec. 2015), 1–17.

[23] Xiong Xiao, Xiaohai Tian, Steven Du, Haihua Xu, Eng Chng, and Haizhou Li. 2015. Spoofing Speech Detection Using High Dimensional Magnitude and Phase Features: the NTU System for ASVspoof 2015 Challenge. In *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. 2052–2056.

[24] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al. 2021. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537*.

[25] Hong Yu, Zheng-Hua Tan, Zhanyu Ma, Rainer Martin, and Jun Guo. 2018. Spoofing Detection in Automatic Speaker Verification Systems Using DNN Classifiers and Dynamic Acoustic Features. *IEEE Trans. Neural. Netw. Learn. Syst.* 29, 10 (Oct. 2018), 4633–4644.