Spoofing-Aware Speaker Verification with Unsupervised Domain Adaptation

Xuechen Liu, Md Sahidullah, Tomi Kinnunen

Spoofing-Aware Speaker Verification with Unsupervised Domain Adaptation

Xuechen Liu[1,2]*, Md Sahidullah*[2]*, Tomi Kinnunen*[1]

[1]School of Computing, University of Eastern Finland, Joensuu, Finland
[2]Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France
xuechen.liu@inria.fr

Abstract

In this paper, we initiate the concern of enhancing the spoofing robustness of the automatic speaker verification (ASV) system, without the primary presence of a separate countermeasure module. We start from the standard ASV framework of the ASVspoof 2019 baseline and approach the problem from the back-end classifier based on probabilistic linear discriminant analysis. We employ three unsupervised domain adaptation techniques to optimize the back-end using the audio data in the training partition of the ASVspoof 2019 dataset. We demonstrate notable improvements on both logical and physical access scenarios, especially on the latter where the system is attacked by replayed audios, with a maximum of 36.1% and 5.3% relative improvement on bonafide and spoofed cases, respectively. We perform additional studies such as per-attack breakdown analysis, data composition, and integration with a countermeasure system at score-level with Gaussian back-end.

Keywords: spoofing-aware speaker verification, anti-spoofing, unsupervised domain adaptation.

1. Introduction

Automatic speaker verification (ASV) [1] is one of the most natural and convenient ways for biometric person authentication. Moving on from conventional models such as *Gaussian mixture models* (GMM) and *i-vector*, state-of-the-art ASV systems based on deep neural networks (DNNs) show reasonably good recognition accuracy on different speech corpora such as NIST SREs, VoxCelebs, and SITW [2]. However, ASV performance is substantially degraded in the presence of spoofing attacks made with voice conversion, text-to-speech synthesis, and replay [3], [4]. High vulnerability of ASV system has been demonstrated using the speech corpora developed for automatic speaker verification spoofing and countermeasures challenge (ASVspoof) [5]–[8].

To protect the ASV systems from spoofing attacks, spoofing countermeasures (CM) are required which can discriminate the natural or human speech from spoofed or computer-generated/playback speech [3], [4]. By leveraging data resources such as ASVspoof challenges, the main focus of the community has been in developing a dedicated CM suitable for detecting a wide variety of spoofing attacks. Then, the separately designed CM is integrated with the ASV system at score-level or decision-level [9]. The work in [7] applied Gaussian back-end based fusion of ASV and CM scores. More recently, joint optimization of ASV and CM was performed on embedding space to detect both human and spoofed imposters [10]. The combined system has shown substantial performance improvement when dealing with imposters using spoofed audio.

While the integrated systems have demonstrated better performance than standalone ASV, such system combination raises several concerns. First, this strategy requires additional workloads for design, training, and optimization. Second, the low generalizability of the CM system can have a severe impact on integrated systems even though ASV systems show relatively better generalization. Moreover, the presently used integration methods have their limitations and they often degrade overall ASV performance by increasing either false acceptance rate or false rejection rate for trials with bonafide speech. Recently the *spoofing-aware speaker verification (SASV) challenge* has been announced [11], aiming at designing spoofing-aware ASV systems. However, the problem that emerged with CM persists.

Inspired by the above concerns, in this work, we initiate the need of making the ASV system itself more aware of spoofing attacks, without the primary presence of CM. We particularly focus on *unsupervised domain adaptation* (DA) by considering the spoofed audio samples available for training CM systems. We hypothesize that the use of spoofed audio data during domain adaptation would provide better generalization for ASV systems to spoofed imposters. To the best of our knowledge, this is the first comparative study on unsupervised DA techniques aiming at improving the generalization power of the ASV system towards the needs of anti-spoofing.

2. Unsupervised Domain Adaptation

One of the key challenges in speaker recognition is generalization across different domains. Whether due to channel, speaker population, language (or other) factors, mismatch in training and test data induces substantial performance drop. While this *domain mismatch* problem can be alleviated in different ways, it is particularly convenient to update the back-end classifier to be better matched with the new domain. In this section, we thus describe the methods used to address the problem of creating spoofing-aware ASV, tackling the back-end system based on *probabilistic linear discriminant analysis* (PLDA) [12], a Gaussian-based classifier that has been widely used for ASV.

2.1. PLDA

PLDA models both channel and speaker variability through specifically structured covariance matrices. By denoting speaker embedding as $\boldsymbol{\phi}$, PLDA models the speaker embedding space as follows [12]–[14]:

$$p(\boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{\phi}|\boldsymbol{\mu}, \boldsymbol{\Phi}_b + \boldsymbol{\Phi}_w), \tag{1}$$

where $\boldsymbol{\mu}$ is the global mean vector, $\boldsymbol{\Phi}_b$ and $\boldsymbol{\Phi}_w$ respectively model the between-class and within-class covariances.

Since we use simplified PLDA [15], by denoting $\mathbf{F}$ and $\mathbf{G}$ as the speaker and channel loading matrices respectively, the two covariance matrices are structured as $\boldsymbol{\Phi}_b = \mathbf{F}\mathbf{F}^T$,

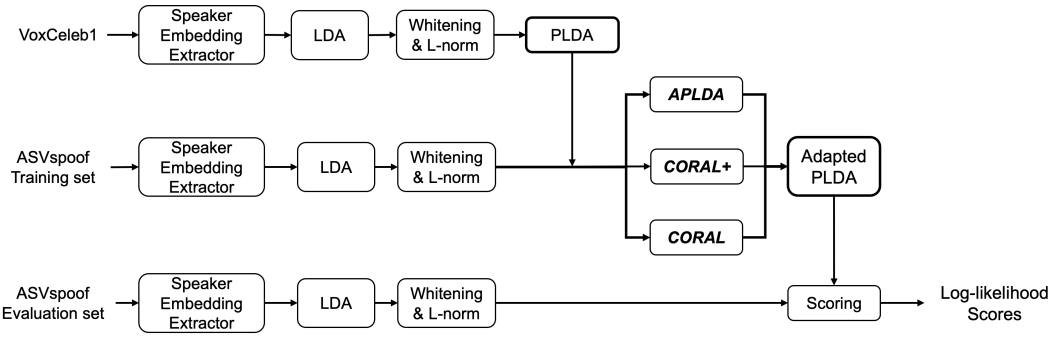

Figure 1: An illustration of the back-end system, along with introduced unsupervised domain adaptation techniques.

$\mathbf{\Phi}_w = \mathbf{G}\mathbf{G}^T + \mathbf{\Sigma}$, where $\mathbf{\Sigma}$ is a diagonal matrix which models residual covariances. For more information about the training and scoring procedure of PLDA, the readers are refer to [13], [14], [16].

2.2. Domain Adaptation with ASVspoof

The problem of domain mismatch remains for PLDA despite its promising performance in general, which needs the presence of in-place training data. Nevertheless, for complicated scenarios it might be hard to obtain enough amount of data to train a PLDA model. Since the quantity of data from the new domain might be limited, a common practice on this issue is to first train a PLDA using out-of-domain (OOD) data and adapt it with a small quantity of in-domain (inD) data from the new domain. Depending on speaker labeling of the in-domain data, both supervised (speaker labels required) [10] and unsupervised (speaker labels not required) [15] methods are available. In this study, we focus on the latter not only because of its broader scope but also from a more fundamental perspective as our scenario involves the use of *spoofed* audio-data in adaptation where speaker labels are not necessarily unambiguously defined. For instance, a VC-transformed speaker may 'blend' voice characteristics of both source and target speakers.

For the adaptation, we implement various methods with the bonafide speaker labels from ASVspoof 2019 in hand. We unravel the efficacy of unsupervised domain adaptation methods with a PLDA trained with OOD data. As for the in-domain (InD) data, we explore with bonafide-only and bonafide-and-spoofed partitions (specified in Section 4.1). Our back-end system with the methods is illustrated in Fig. 1. The unsupervised DA methods covered in this paper are discussed below.

Correlation Alignment (CORAL). CORAL [17] is a feature-based DA technique that aligns the covariance matrices of OOD features to compensate the InD ones by minimizing the distance between them. It aims at transforming the embedding space of the original data to the target one while preserving maximal similarity to the former. Each embedding is transformed as

$$\boldsymbol{\phi} \leftarrow \mathbf{C}_{\mathrm{I}}^{\frac{1}{2}} \mathbf{C}_{\mathrm{o}}^{-\frac{1}{2}} \boldsymbol{\phi}, \tag{2}$$

where $\mathbf{C}_{\mathrm{I}}$ and $\mathbf{C}_{\mathrm{o}}$ are the covariance matrices computed from InD and OOD data respectively. The transformation is done via *zero-phase component analysis* (ZCA) [18]. The transformed pseudo-InD embeddings are used to re-train the PLDA. CORAL has been applied to speaker verification in [13], [15], [19].

CORAL+. Alternating from CORAL, CORAL+ [13] is a model-based method working on covariance matrices of the PLDA directly. Instead of replacing the covariance matrices $\mathbf{\Phi}_{\mathrm{b}}$ and $\mathbf{\Phi}_{\mathrm{w}}$ with ones computed from the transformed pseudo-InD data, CORAL+ updates them by linearly interpolating between the original and the pseudo-InD ones:

$$\begin{aligned} \mathbf{\Phi}_b &\leftarrow (1-\beta)\mathbf{\Phi}_b + \beta \mathbf{A}^T \mathbf{\Phi}_b \mathbf{A} \\ \mathbf{\Phi}_w &\leftarrow (1-\lambda)\mathbf{\Phi}_w + \lambda \mathbf{A}^T \mathbf{\Phi}_w \mathbf{A} \end{aligned} \tag{3}$$

with control parameters $\beta, \lambda \in [0, 1]$. $\mathbf{A} = \mathbf{C}_{\mathrm{I}}^{\frac{1}{2}} \mathbf{C}_{\mathrm{o}}^{-\frac{1}{2}}$ is the transformation matrix. In order to enhance the uncertainty accounted and cover larger speaker subspace, CORAL+ performs *simultaneous diagonalization* [20] on both the original and pseudo-InD covariances. Details of CORAL+ can be found in [13]. CORAL+ has been reported to outperform CORAL in DNN-based ASV [13], [15].

Kaldi adaptation (APLDA). The unsupervised PLDA adaptor from Kaldi toolkit [15] starts by performing eigen decomposition on the total covariance matrices $\mathbf{\Sigma}_0^{\frac{1}{2}} \mathbf{\Sigma}_i \mathbf{\Sigma}_0^{\frac{1}{2}} = \mathbf{P}\mathbf{\Delta}\mathbf{P}^T$, where $\mathbf{\Sigma}_0 = \mathbf{\Phi}_{\mathrm{b,0}} + \mathbf{\Phi}_{\mathrm{w,0}}$ and $\mathbf{\Sigma}_i = \mathbf{\Phi}_{\mathrm{b,i}} + \mathbf{\Phi}_{\mathrm{w,i}}$ are covariance matrices computed from inD and inD and OOD data respectively. The obtained eigenvalues (as diagonal matrix $\mathbf{\Delta}$) and vectors (as matrix $\mathbf{P}$) are then used to update the PLDA covariances, followed by simultaneous diagonalization. It shares operations in common with CORAL and CORAL+, except that it performs interpolation on diagonal values of the PLDA parameters via eigenvalues, instead of on the parameters themselves. Mathematical details and its application for DNN-based ASV can be found in [15]. This is the method used in ASV system developed for ASVspoof 2019 challenge [21].

3. Dataset Description: ASVspoof 2019

We conduct the experiments on ASVspoof 2019 corpus [22]. Originally launched in 2015, the biennial ASVspoof challenge[1] series focus on assessing the vulnerability of ASV systems against different spoofing attacks and developing standalone countermeasures. The ASVspoof 2019 database furthers the achievements and protocol from its 2015 and 2017 predecessors, with a more controlled setup and evaluation protocols. It acquires more state-of-the-art neural-based algorithms on TTS and VC, as well as more careful simulation of replayed speech,

[1] https://www.asvspoof.org/

which makes it also useful for fraud audio detection in real-time cases such as telebanking and smart homes.

The database has two subsets: *logical access* (LA) and *physical access* (PA). The LA subset corresponds to a scenario where the attacks come in the form of synthetic and converted speech. Such speech cannot be perceptually detected by humans but can be distinguished by ASV systems if equipped with reliable models. State-of-the-art TTS algorithms were applied to construct the database, including but not limited to variational autoencoder (VAE) [23], WaveCycleGAN [24], and Tacotron [25]. They were equipped by advanced vocoders such as WaveNet [26] and WORLD [27], generating high-quality synthetic speech. The vocoders mentioned here were also used for VC. The PA subset corresponds to the case where the attacks are presented in a simulated physical space with varying positions of the speaker, microphone, and reverberation time. In the evaluation, two main factors are considered: distance between attacker and the speaker and the quality of the replaying device.

Both the subsets have their dedicated training, development, and evaluation partition which are commonly used for assessing spoofing countermeasures. Apart from these, the dataset comes with separate enrollment files for ASV experiments. The audio data provided for training CM systems could be employed for ASV domain adaption. The summary of the protocols of the two subsets in terms of the number of trials is shown in Table 1.

Trial Type	LA		PA	
	dev	eval	dev	eval
target	1484	5370	2700	12960
bonafide non-target	5768	33327	14040	123930
spoofed non-target	22296	63882	24300	116640

Table 1: Trial statistics for ASVspoof 2019.

4. Experiments

4.1. Data

Our speaker embedding extractor is trained on the pooled training sets of VoxCeleb1 [28] and VoxCeleb2 [29] consisting of 7205 speakers. The OOD PLDA is trained on VoxCeleb1, with 1211 speakers.

We use the various partitions of ASVspoof 2019 for domain adaptation. The bonafide training partition is the same bonafide data used for countermeasure training in [21]. It contains 2580 and 5400 utterances for the LA and PA scenarios, respectively. The number of speakers is 20 in both scenarios. We use two subsets correspondingly: 1) *bonafide*, which contains bonafide human speech only; 2) *spoofed*, which is a collection of spoofed utterances corresponding to the same 20 speakers. The amount of data for the latter subset is n-times more than the former, where n is the number of spoofing conditions.

4.2. System Configuration

In all experiments, we use 40-dimensional mel filterbanks with Hamming window as the acoustic features. The size and step size of the Hamming window is 25ms and 10ms, respectively, and the number of FFT bins is 512. We use *x-vector* [30] based on *extended time-delayed neural network* (E-TDNN) [31]. Differently from [31], we replace the statistics pooling layer with attentive statistics pooling [32] and employ *additive angular softmax* [33] as the training loss. We extract the embedding for each input utterance from the first fully-connected layer after the pooling layer.

The extracted vectors are centered, unit-length normalized, and projected with a 150-dimensional LDA, to train and adapt the PLDA. We adapt the OOD PLDA with the aforementioned DA methods, with the scaling factor of within-class and between-class covariances being set to be different for the two scenarios, following the protocol described in [21]: $\alpha_{\mathrm{w}} = 0.25$, $\alpha_{\mathrm{b}} = 0.0$ for LA, and $\alpha_{\mathrm{w}} = 0.9$, $\alpha_{\mathrm{b}} = 0.0$ for PA.

4.3. Evaluation

We create trials for both the LA and PA scenarios following [21]. Additionally, we utilize two trials lists in each scenario: *bonafide* trials includes a mix of bonafide target and zero-effort impostor trials", while *spoofed* trials are composed by bonafide target and spoofed trial pairs, treated as impostors. Log-likelihood ratio (LLR) scores are produced for the trials and *equal error rate* (EER) is used to gauge performance. As some of our analyses report EERs on per-attack bases, with a limited number of trials, we also report the parametric 95% confidence intervals $(EER \pm \delta * Z_{\alpha/2}) * 100\%$, following methods described in [34]. We set the related parameters $\delta = 0.5\sqrt{\mathrm{EER}(1-\mathrm{EER})n_+ + n_-)/(n_+ * n_-)}$ and $Z_{\alpha/2} = 1.96$, where n_+ and n_- are the number of target and nontarget trials, respectively.

		ASV EER(%)	
Method	Data for adapt.	*bonafide*	*spoofed*
–	–	1.43	**35.55**
CORAL	*bonafide*	**1.04**	37.32
CORAL	*bonafide+spoof*	1.09	36.01
CORAL+	*bonafide*	1.11	37.44
CORAL+	*bonafide+spoof*	1.17	35.94
APLDA	*bonafide*	1.14	37.06
APLDA	*bonafide+spoof*	1.17	35.61

Table 2: Results on ASVspoof 2019 LA evaluation.

		ASV EER(%)	
Method	Data for adapt.	*bonafide*	*spoofed*
–	–	5.46	39.95
CORAL	*bonafide*	3.75	39.48
CORAL	*bonafide+spoof*	**3.49**	39.07
CORAL+	*bonafide*	3.86	39.41
CORAL+	*bonafide+spoof*	3.61	**37.85**
APLDA	*bonafide*	4.11	39.79
APLDA	*bonafide+spoof*	3.79	39.27

Table 3: Results on ASVspoof 2019 PA evaluation.

5. Results and Analysis

5.1. Logical Access

Table 2 presents the results for the LA scenario. We first compare different methods with bonafide data for adaptation. In terms of bonafide EER, all the methods outperform the baseline (no adaptation), as expected. The maximum relative improvement of 27.3% is provided by CORAL. Meanwhile, on EER from spoofed trials, the baseline achieves the lowest EER across all systems. APLDA with bonafide and spoof audios for

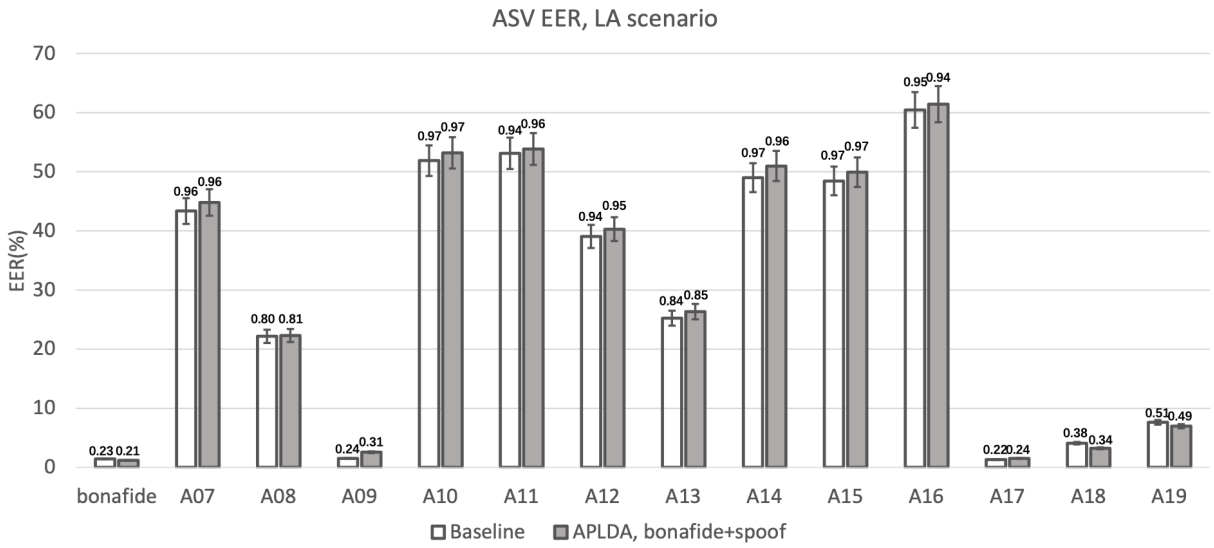

Figure 2: Breakdown of ASV EER with baseline and best-performed method with respect to spoofing attacks for LA scenario. Numbers in small font size are the parametric confidence intervals.

		ASV EER(%)	
Method	Sampling	*bonafide*	*spoofed*
APLDA	*per-spk*	1.12	37.17
APLDA	*per-attack*	1.17	37.13
APLDA	*per-both*	1.17	35.67

Table 4: LA results of ablation study on data composition.

		ASV EER(%)	
Method	Sampling	*bonafide*	*spoofed*
CORAL+	*per-spk*	4.07	39.31
CORAL+	*per-attack*	4.11	39.25
CORAL+	*per-both*	3.64	38.84

Table 5: PA results of ablation study on data composition.

adaptation reaches more comparable performance against the baseline.

We then focus on the effect of spoofed data for the adaptation. While including spoofed audios for adaptation increases bonafide EER across all methods, there are improvements in EER for spoofed trials. Maximum relative improvement on spoofed EER comes from CORAL+, by 4.1%. Nevertheless, the increased amount of data does not outperform the baseline for the DA methods. Noting that LA corresponds to the cases where synthesized and converted speech segments are used for attacking, the results suggest that unsupervised DA methods may not be effective on generalization for synthesized spoofing attacks.

A breakdown of the results per attack is shown in Fig. 2 for the best-performing system (APLDA). The baseline provides slightly lower EERs (though without significant differences) for most attacks. On attacks A18 and A19, APLDA outperforms baseline. Referring to [21], in these two attacks the spoofed audio is generated with VC via transfer learning from conventional statistical models such as GMM and i-vector, while other types of attack are mostly constructed via neural-based approaches. Similar observations can be found for other methods not shown in the figure as well.

5.2. Physical Access

Table 3 presents the results for the PA scenario. Focusing on systems with bonafide data for adaptation, notable improvements over the baseline are obtained. The maximum relative reduction on bonafide EER is 31.3%, provided by CORAL. CORAL+ gives the lowest spoofed EER, lower than the baseline by relatively 1.4%.

Different from LA, augmenting adaptation data by spoofed utterances further improves the performance of all three DA methods. The lowest bonafide EER is achieved by CORAL, with 36.1% relative reduction over the baseline and 6.9% relative reduction over the corresponding bonafide-only adaptation. Similar to LA, CORAL+ is efficient on spoofed evaluation trials, with 5.3% relative EER reduction over the baseline. These observations suggest the potential of unsupervised adaptation on increasing awareness of the ASV system against computer-simulated room replay attacks.

A break-down of the results per attack is shown in Fig. 3 for CORAL+. Different from LA, there are more attacks where the DA method outperforms the baseline. Recall from [21] that each attack is made of a duple, where the first letter refers to the distance between the attacker and the speaker (‘A’ is closest, ‘C’ is farthermost) and the second letter refers to the quality of replay device (‘A’ denotes the best, ‘C’ denotes the worst). The lowest EERs are obtained for attacks starting with ‘B’ and ‘C’, where the distance between the attacker and the talker is larger. In relative terms, these attacks are observed as being less detrimental. This indicates the potential of DA methods in handling attacks from more distant positions, while rather less affected by the quality of the replay device.

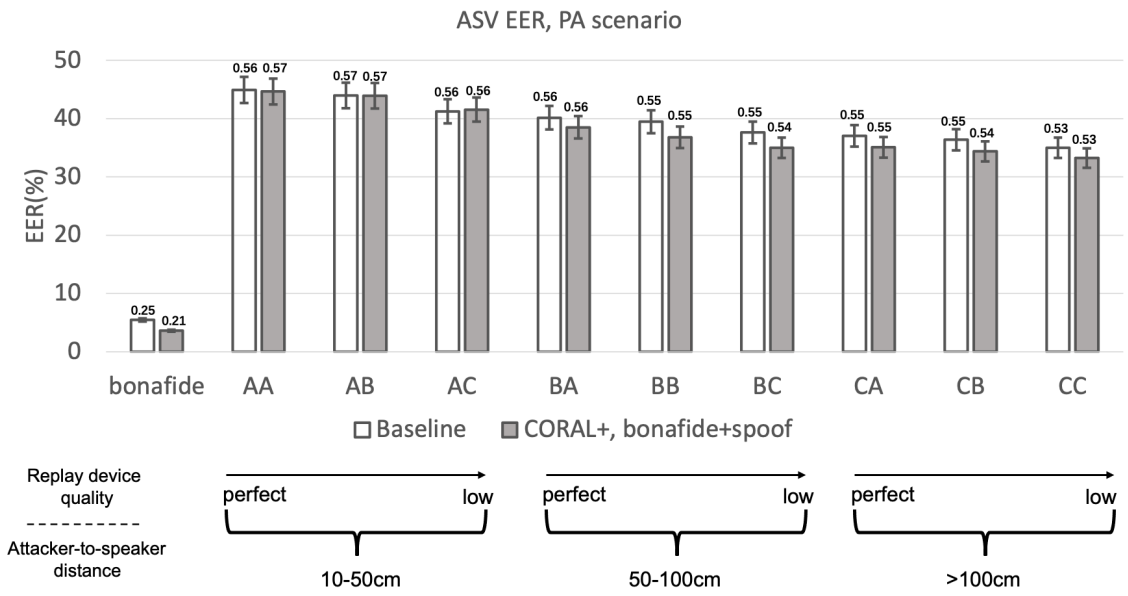

Figure 3: Breakdown of ASV EER with baseline and best-performed method with respect to spoofing attacks for PA scenario, along with the annotation of related variables [21]. Numbers in small font size are the parametric confidence intervals.

5.3. Ablation Study on Data Composition

Up to this point, we have compared different adaptation methods under two sets of adaptation data: bonafide only, and bonafide with spoofed data of bonafide speakers.

A natural question on the amount of data arises: in the above settings, the amount of data between *bonafide* and *bonafide+spoof* are different. This is because the spoofing data are sampled separately and thus contain n times more utterances than the bonafide data, where n denotes the number of attacks present in the spoofed CM training set (6 for LA, 10 for PA). Thus, it is difficult to pinpoint whether performance differences are due to merely an increased amount of data, or rather, the type of data.

To address this question, we conduct an ablation experiment in both scenarios, where the total amount of bonafide and spoof data is held fixed and balanced (1290 utterances for both) but where we vary the composition of the spoof data. To compare with bonafide-only data setup, where assuming we have m utterances, we sample $m/2$ utterances from the bonafide and spoofed set. For the bonafide part, we sample with the same size, with all speakers (20 of them) kept. We consider three alternative approaches:

- *per-spk*: Sampling per speaker, where we sample $s_1 = m/(2 \times 20)$ utterance for each speaker from the spoofed set;
- *per-attack*: Sampling per attack, where we sample $s_2 = m/(2 \times n)$ utterances from the spoofed set for each attack;
- *per-both*: The intersection of both the two factors. The total number of conditions here is $s_1 \times s_2$ and we sample $n/(s_1 \times s_2)$ utterances from each condition.

We separate the discussion concerning different scenarios and choose the best-performing system in terms of ASV EER on *spoofed* trials (APLDA for LA, and CORAL+ for PA).

		ASV+CM EER(%)	
Method	Data for adapt.	*bonafide*	*spoofed*
–	–	1.72	0.97
CORAL	*bonafide*	1.69	0.95
CORAL	*bonafide+spoof*	1.69	0.96
CORAL+	*bonafide*	1.73	0.96
CORAL+	*bonafide+spoof*	1.75	0.96
APLDA	*bonafide*	1.70	0.97
APLDA	*bonafide+spoof*	1.81	0.99

Table 6: Results on LA, with integrated CM.

The results are presented in Tables 4 and 5. For LA, compared to adaptation with only bonafide data, *per-both* leads to the most competitive spoofed performance, with a relatively 18.1% lower EER. This highlights the positive effect of spoofed data on enhancing the robustness of the system. Meanwhile, more spoofed data does not return significantly better performance on both trials, which indicates the relatively-low effect of the amount of spoofed data on improving the awareness of the system on synthetic speech.

For PA which corresponds to the replay attacks, replacing part of bonafide speech with spoofed ones does not lead to significantly better performance even on bonafide-only evaluation. The best one across different sampling schemes comes from the one covering all speakers and types of attack, outperforming adaptation with only bonafide data by relatively 1.4%. The increasing amount of spoofed data this time leads to better performance, which is a different observation from LA.

5.4. Integration with Countermeasure System

Up to this point, we have compared different adaptation methods in terms of their efficacy on a standalone ASV system. We would then like to investigate its integration with a fixed CM system.

		ASV+CM EER(%)	
Method	Data for adapt.	*bonafide*	*spoofed*
–	–	8.77	2.73
CORAL	*bonafide*	7.89	2.77
CORAL	*bonafide+spoof*	7.74	2.76
CORAL+	*bonafide*	7.92	2.80
CORAL+	*bonafide+spoof*	7.74	2.69
APLDA	*bonafide*	8.27	2.81
APLDA	*bonafide+spoof*	7.97	2.71

Table 7: Results on PA, with integrated CM.

We perform the parallel integration at score level using the Gaussian back-end fusion method described in [35]. It utilizes 2-dimensional vector $\mathbf{s} = [s_{\text{CM}}, s_{\text{ASV}}]^T$ consisting of CM score s_{CM} and ASV score s_{ASV}. We define three classes: target, zero-effort impostor (aka non-target), and spoof impostor. Each of these three classes is modeled with a bivariate Gaussian. The 2D mean vectors and 2×2 full covariance matrices of each class are obtained using their maximum likelihood estimates. Log-likelihood ratio scores for new trials are then computed by treating targets as the positive class (numerator) and the combined class of zero-effort impostors and spoof impostors as the negative class (denominator). Note that the latter is a two-component Gaussian mixture distribution. The mixing weight of each mixture component is set as $\alpha = 0.5$. While the ASV scores correspond to one of the systems described above, the CM scores are produced using the method in [36]. We train *light convolutional network* (LCNN) models for the LA and the PA scenarios using the respective training sets of ASVspoof 2019.

The fusion results are displayed in Table 6 and 7 for the LA and PA scenarios, respectively. As expected, fusion boosts the performance dramatically. On the LA scenario, the difference between the baseline and the DA methods reduces, regardless of the DA method or adaptation data. Similar observations can be found for PA, where the performance gap on the spoofed set is larger compared with LA, although slight degradation can be meanwhile observed for both scenarios. CORAL+ with both types of data for adaptation maintains the best performance on both bonafide and the spoofed evaluation sets across all systems. However, its relative improvement against the baseline is narrow. These findings indicate the gap between adaptation methods and separate CM ingredients, and the investigation of the former as open questions.

5.5. Comparison with recent SASV challenge baseline

Finally, we compare our results from the LA scenario with the (currently available) ASV results on the ongoing SASV challenge [37] (there is no PA scenario in this challenge) which shares the same data and protocols adopted for this study. Despite the shared evaluation setup, the compared ASV systems are very different. Thus, our aim is not a detailed discussion of the advantages or disadvantages of each architecture but, rather, reassurance that our results are reasonably well aligned with the (currently-reported) results of the SASV challenge.

This comparison is shown in Table 8. Both our baseline and selected system (that reach the best performance on spoofed trial) return comparable performance against the SASV, with a relative improvement of 28.7% in terms of bonafide EER provided by APLDA with bonafide and spoofed data for the adaptation.

	ASV EER(%)	
System	*bonafide*	*spoofed*
ECAPA-TDNN [37]	1.64	30.75
Baseline, no DA	1.43	35.55
APLDA, *bonafide+spoof*	1.17	35.61

Table 8: Comparison with SASV baseline.

6. Conclusion

We have conducted a preliminary study on spoofing-aware ASV with a special focus on unsupervised domain adaptation of a PLDA back-end. While our experiments also address fusion of ASV with a standalone spoofing countermeasure, our work is substantially differentiated from the majority of work on anti-spoofing that focuses on improving standalone countermeasures.

The key benefit of our unsupervised approach is simplicity. As the supply of spoofed speech data resources keeps increasing year by year, and since no architectural modifications to a conventional speaker recognition system is needed, it is straightforward to apply the technique to update existing PLDA back-end models for scenarios demanding increased spoof-awareness.

While improvements on both bonafide and spoofed trials were obtained (especially in the PA scenario) through unsupervised domain adaptation, it is also evident that the absolute error rates on the spoofed trials remain too high on spoofing attacks. This may suggest that it is challenging to make a conventional speaker embedding extractor with PLDA back-end work on a mix of bonafide and spoofed data. Given the related activities in the ongoing SASV challenge [37], we have have to reconsider either entirely different speaker embeddings, back-ends — or both.

7. References

[1] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[2] Z. Bai and X. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[3] Z. Wu, N. Evans, T. Kinnunen, *et al.*, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[4] M. Sahidullah, H. Delgado, M. Todisco, *et al.*, "Introduction to voice presentation attack detection and recent advances," in *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection*, S. Marcel, M. S. Nixon, J. Fierrez, *et al.*, Eds. 2019, pp. 321–361.

[5] Z. Wu, T. Kinnunen, N. Evans, *et al.*, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Interspeech*, 2015, pp. 2037–2041.

[6] T. Kinnunen, M. Sahidullah, H. Delgado, *et al.*, "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection," in *Proc. Interspeech*, 2017, pp. 2–6.

[7] M. Todisco, X. Wang, V. Vestman, *et al.*, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *Proc. Interspeech*, 2019, pp. 1008–1012.

[8] J. Yamagishi, X. Wang, M. Todisco, *et al.*, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 47–54.

[9] M. Sahidullah, H. Delgado, M. Todisco, *et al.*, "Integrated Spoofing Countermeasures and Automatic Speaker Verification: An Evaluation on ASVspoof 2015," in *Proc. Interspeech*, 2016, pp. 1700–1704.

[10] A. Gomez-Alanis, J. A. Gonzalez-Lopez, S. P. Dubagunta, *et al.*, "On joint optimization of automatic speaker verification and anti-spoofing in the embedding space," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1579–1593, 2021.

[11] J.-w. Jung, H. Tak, H.-j. Shim, *et al.*, "SASV challenge 2022: A spoofing aware speaker verification challenge evaluation plan," *CoRR*, vol. abs/2201.10283, 2022.

[12] S. Ioffe, "Probabilistic linear discriminant analysis," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 531–542.

[13] K. A. Lee, Q. Wang, and T. Koshinaka, "The CORAL+ algorithm for unsupervised domain adaptation of plda," in *Proc. ICASSP*, 2019, pp. 5821–5825.

[14] A. Sizov, K. A. Lee, and T. Kinnunen, "Unifying probabilistic linear discriminant analysis variants in biometric authentication," in *Structural, Syntactic, and Statistical Pattern Recognition*, Springer Berlin Heidelberg, 2014, pp. 464–475.

[15] P.-M. Bousquet and M. Rouvier, "On Robustness of Unsupervised Domain Adaptation for Speaker Recognition," in *Proc. Interspeech*, 2019, pp. 2958–2962.

[16] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–8.

[17] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16, Phoenix, Arizona: AAAI Press, 2016, pp. 2058–2065.

[18] A. Kessy, A. Lewin, and K. Strimmer, "Optimal whitening and decorrelation," *The American Statistician*, pp. 1–6, Dec. 2016.

[19] M. J. Alam, G. Bhattacharya, and P. Kenny, "Speaker verification in mismatched conditions with frustratingly easy domain adaptation," in *Proc. Speaker Odyssey*, 2018, pp. 176–180.

[20] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1990.

[21] A. Nautsch, X. Wang, N. Evans, *et al.*, "Asvspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.

[22] X. Wang, J. Yamagishi, M. Todisco, *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101 114, 2020.

[23] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada*, 2014.

[24] K. Tanaka, T. Kaneko, N. Hojo, *et al.*, "Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 632–639.

[25] G. Sanjay, K. C. Sooraj, and D. Mishra, "Natural text-to-speech synthesis by conditioning spectrogram predictions from transformer network on waveglow vocoder," in *2020 7th International Conference on Soft Computing Machine Intelligence (ISCMI)*, 2020, pp. 255–259.

[26] A. van den Oord, S. Dieleman, H. Zen, *et al.*, "WaveNet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[27] M. MORISE, F. YOKOMORI, and K. OZAWA, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[28] A. Nagrani, J. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[29] J. Chung *et al.*, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[30] D. Snyder, D. Garcia-Romero, G. Sell, *et al.*, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[31] D. Snyder, D. Garcia-Romero, G. Sell, *et al.*, "Speaker recognition for multi-speaker conversations using x-vectors," in *Proc. ICASSP*, 2019, pp. 5796–5800.

[32] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Proc. Interspeech*, 2018, pp. 2252–2256.

[33] F. Wang, J. Cheng, W. Liu, *et al.*, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[34] S. Bengio and J. M. thoz, "A statistical significance test for person authentication," in *Proc. The Speaker and Language Recognition Workshop*, 2004, pp. 176–180.

[35] M. Todisco, H. Delgado, K. A. Lee, *et al.*, "Integrated presentation attack detection and automatic speaker verification: Common features and gaussian back-end fusion," in *Proc. Interspeech*, 2018, pp. 77–81.

[36] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," in *Proc. Interspeech*, 2021, pp. 4259–4263.

[37] Y. Zhang, G. Zhu, and Z. Duan, "A new fusion strategy for spoofing aware speaker verification," *CoRR*, vol. abs/2202.05253, 2022.