Diagnosing and Treating Audio-Video Fake Detection

Marcel Klemt Carlotta Segna Anna Rohrbach
Technische Universität Darmstadt, hessian.AI
{marcel.klemt, carlotta.segna, anna.rohrbach}@tu-darmstadt.de

Abstract

Generative AI advances rapidly, allowing the creation of realistic video and audio. This progress presents a significant security/ethical threat, as malicious users can exploit DeepFake techniques to spread misinformation. Recent DeepFake detection approaches explore the multimodal (audio-video) threat scenario. Still, there are challenges hindering progress, in part caused by limited reproducibility and many issues with existing datasets, including the discovery of a leading silence shortcut in the widely used FakeAVCeleb dataset (and possibly others). We address these issues one step at a time. We propose a SImple Multimodal BAseline (SIMBA), achieving performance comparable to SoTA models while maintaining a minimalistic design. We spotlight the recent DeepSpeak v1 dataset, being the first to propose an evaluation protocol and benchmark it with SoTA models. Further, we analyze the FakeAVCeleb dataset, uncovering blind spots in the prior evaluation and proposing a revised evaluation protocol. Finally, we contribute an augmentation scheme to tackle the leading silence shortcut. Our findings offer a way forward in the important area of audio-video DeepFake detection.

1. Introduction

Nowadays, Internet users can create, thanks to ready-to-be-used applications, new and extremely realistic DeepFakes of friends, politicians, and strangers with just a few seconds of footage. This creates a strong need to address the threat posed by DeepFake technology, as it can be easily used for malicious purposes. To counter the DeepFake threat, early work primarily focused on unimodal scenarios, e.g., video-only approaches [8, 7]. Recently, multimodal audio-visual DeepFake detection methods are becoming more prominent [3, 4, 15, 13, 6]. Unfortunately, many approaches do not have code available.

In order to train multimodal audio-video DeepFake detection models, several datasets have been proposed [5, 10, 9, 1, 15, 11]. Yet, most of these datasets suffer from one or more issues. (a) Some datasets, e.g., KoDF [10] and AVLips [11], only offer manipulations for the visual modality, preventing the study of diverse combinations of manipulated modalities. (b) Datasets like DFDC [5] and AVLips [11] only provide binary labels, thus not supporting the evaluation of cross-manipulation generalization, which is important for practical applicability. (c) The recent work [2] disclosed the presence of shortcuts, e.g., in FakeAVCeleb [9] manifested as leading silence which gives away some manipulations. Shortcuts undermine dataset utility, as performance can no longer be trusted. (d) Some popular datasets are fairly saturated [13, 8, 7]. To allow further successful development of audio-visual DeepFake detection models and measure the progress, new realistic benchmarks and thorough evaluation protocols are required.

To address these issues, we highlight the very recent and previously unexplored dataset DeepSpeak v1 [1] and make a case for its use as a new benchmark for multimodal DeepFake detection. DeepSpeak v1 contains more recent manipulation techniques, extreme head poses, and occlusions. We are, to the best of our knowledge, the first to present an evaluation protocol for DeepSpeak v1 and benchmark SoTA models against it. Our new protocol allows a cross-manipulation evaluation, consisting of several leave-one-out *method* splits and *family* splits, which present generalization tasks of varying difficulty. To be able to analyze the dataset with a supervised multimodal DeepFake detection model, we introduce a new SImple Multimodal BAseline (SIMBA). SIMBA stands out with its simple design yet competitive performance compared to more complex architectures. We explore several augmentation/sampling strategies to make the method robust to the leading silence shortcut. Lastly, we revisit the popular FakeAVCeleb dataset. We uncover *blind spots* and *leakage* in the commonly used cross-manipulation evaluation protocol due to being introduced [6] for self-supervised scenario *distinct from the supervised cross-manipulation generalization*. We propose a new evaluation protocol (similar to the one for DeepSpeak v1) and urge the community to adopt it. Following [2], we confirm that FakeAVCeleb has a shortcut and our proposed augmentation technique eliminates its impact. Our work offers several practical findings that can inform future work and contributes an effective DeepFake detection baseline.

2. SIMBA: SImple Multimodal BAseline

The architecture of SIMBA, our SImple Multimodal BAseline, can be found in Fig. 1. It mainly consists of a video encoder, an audio encoder, and a fusion layer that joins the video and audio representations into a shared feature space. We employ the R(2D+1) [14] architecture as

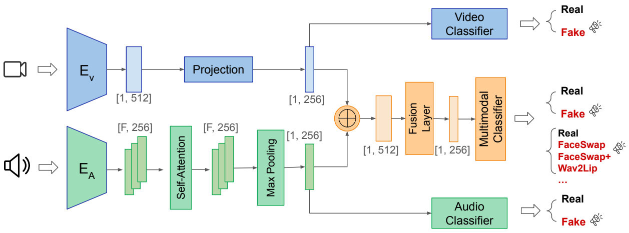


Figure 1: **SIMBA** is composed of an audio and a video branch. Modality features are concatenated ⊕, followed by a fusion layer and a multimodal classifier. We show both the binary and a multiclass variant (the middle branch).

our video encoder backbone and the BYOL-A model [12] architecture as our audio feature extractor. A simple binary classification head is added to each unimodal branch. Video and audio features are given to a fusion layer, which is followed by a multimodal classification head. The multimodal classification head can be trained either in a simple binary *real/fake* fashion or in a multiclass way where the prediction is the *type of manipulation* applied to the input video. Binary Cross-Entropy (BCE) loss is applied for the first setting, and Cross-Entropy (CE) loss in the second setting. The final loss is the sum of the multimodal and both unimodal BCE losses. During inference, the multiclass predictions are transformed into a binary prediction score by summing all predicted fake probabilities (i.e., all but the real class).

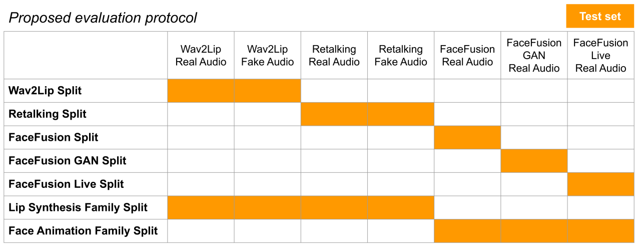
To enhance the robustness of our model to shortcuts, we consider several strategies. Specifically, we employ *temporal jittering* with *consecutive* vs. *subsampled frames*. In the first case, we employ N consecutive frames; for the latter, we sample N frames with a step size of M. Temporal jittering means that we start sampling from an arbitrary position of a training video, *augmenting* the training data.

	FakeAVCeleb	**DeepSpeak**
Binary consecutive	90.39 (-10.89)	91.89 (-5.27)
Multiclass consecutive	95.24 (-15.76)	89.97 (-3.00)
Binary Subsampling jit	89.94 (-0.54)	93.59 (+0.56)
Multiclass Subsampling jit	95.34 (-0.34)	93.06 (+0.13)

Table 1: Results on the FakeAVCeleb and DeepSpeak with untrimmed and trimmed (in parentheses) videos.

3. Dataset Examination

DeepSpeak v1. The DeepSpeak v1 [1] dataset includes 13k videos from 220 identities with fake samples created using five different video and one audio generation technique. To the best of our knowledge, we are the first to propose an evaluation schema for this dataset and benchmark SoTA models on it. Since prior work discovered an audio silence shortcut in existing benchmarks [2], we analyze DeepSpeak v1 in the same way, and find that it also suffers from the shortcut (although to a smaller degree).

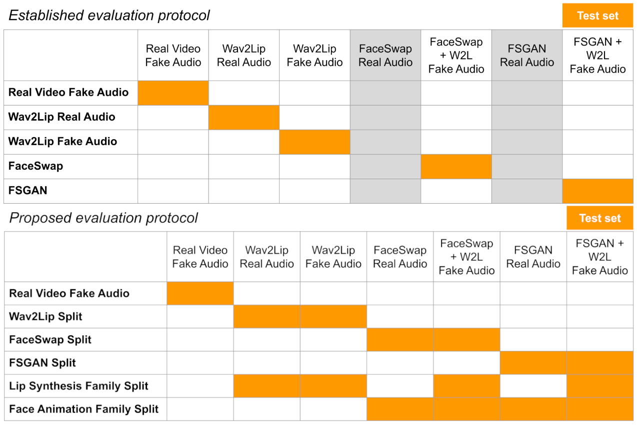


Proposed evaluation protocol

■ Test set

	Wav2Lip Real Audio	Wav2Lip Fake Audio	Retalking Real Audio	Retalking Fake Audio	FaceFusion Real Audio	FaceFusion GAN Real Audio	FaceFusion Live Real Audio
Wav2Lip Split	■	■					
Retalking Split			■	■			
FaceFusion Split					■		
FaceFusion GAN Split						■	
FaceFusion Live Split							■
Lip Synthesis Family Split	■	■	■	■			
Face Animation Family Split					■	■	■

Figure 2: Visualization of the proposed evaluation protocol for DeepSpeak v1 [1]. The first five rows show method splits, whereas the last two rows specify the family splits.



Established evaluation protocol

■ Test set

	Real Video Fake Audio	Wav2Lip Real Audio	Wav2Lip Fake Audio	FaceSwap Real Audio	FaceSwap + W2L Fake Audio	FSGAN Real Audio	FSGAN + W2L Fake Audio
Real Video Fake Audio	■			▨		▨	
Wav2Lip Real Audio		■		▨		▨	
Wav2Lip Fake Audio			■	▨		▨	
FaceSwap				▨	■	▨	
FSGAN				▨		▨	■

Proposed evaluation protocol

■ Test set

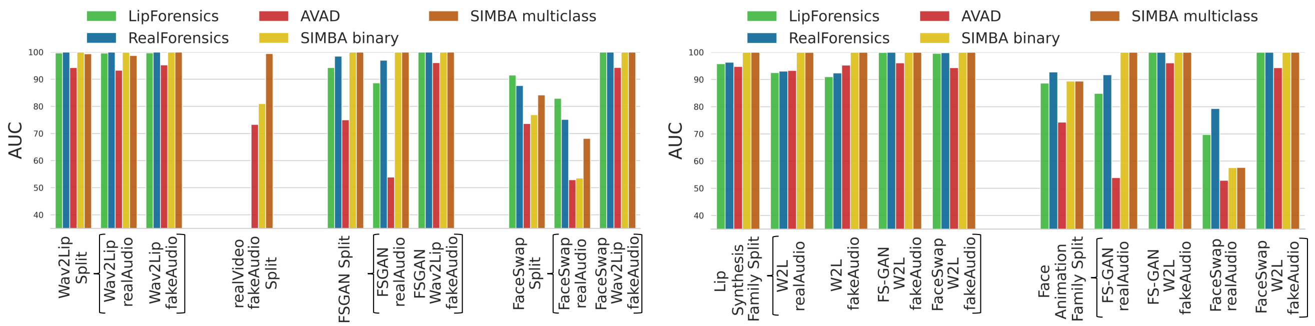
	Real Video Fake Audio	Wav2Lip Real Audio	Wav2Lip Fake Audio	FaceSwap Real Audio	FaceSwap + W2L Fake Audio	FSGAN Real Audio	FSGAN + W2L Fake Audio
Real Video Fake Audio	■						
Wav2Lip Split		■	■				
FaceSwap Split				■	■		
FSGAN Split						■	■
Lip Synthesis Family Split		■	■		■		■
Face Animation Family Split				■	■	■	■

Figure 3: Visualization of the established (top) and our proposed (bottom) cross-manipulation generalization evaluation for FakeAVCeleb [9] (top four rows correspond to *method* splits, the last two rows show the *family* splits).

We propose an evaluation protocol to benchmark SOTA models for a leave-one-out generalization task. Our protocol is divided into two categories: "method" and "family". Each "method" split consists of one type of manipulation with real or fake audio as depicted in Fig. 2 (top five rows). The "family" splits divide the data into the Lip Synthesis vs. Face Animation Family (Fig. 2, last two rows). We prevent leakage of similar artifacts between manipulations and make generalization even harder with our family splits.

FakeAVCeleb. The FakeAVCeleb dataset [9] contains 21k videos from 500 identities and utilizes three video manipulation techniques and one audio manipulation technique. We extend the previous study of the silence shortcut [2] beyond the fake-video-fake-audio samples and find that *leading silence is present in all manipulations with fake audio*. At first, we confirm that SIMBA latches on this artifact similar to other multimodal supervised models [2], but by employing the sampling strategies previously presented, SIMBA becomes robust to the shortcut (see Sec.4).

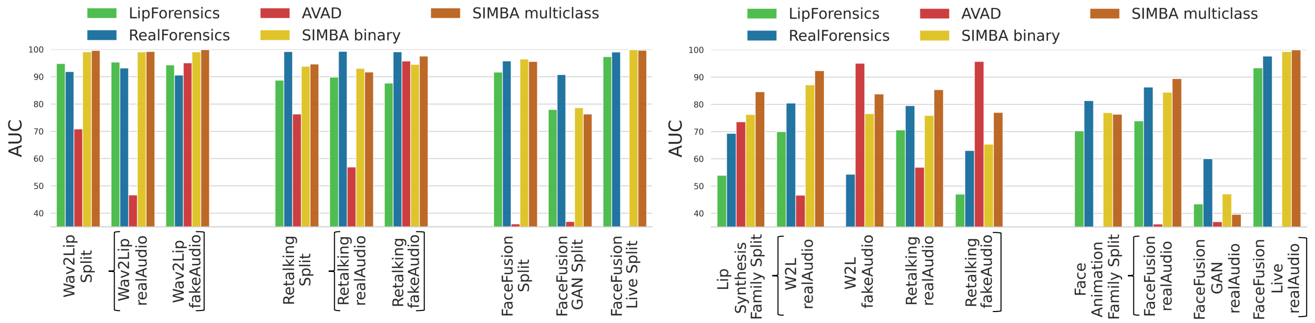
Since FakeAVCeleb is also fairly saturated, we further examine its evaluation protocol. We discover: (a) "leakage" present in the established evaluation scheme (Fig. 3,



(a) FakeAVCeleb Method Split

(b) FakeAVCeleb Family Split

Figure 4: Cross-manipulation comparison in the Method- and Family-based Splits on FakeAVCeleb [9].



(a) DeepSpeak v1 Method Split

(b) DeepSpeak v1 Family Split

Figure 5: Cross-manipulation comparison in the Method- and Family-based Splits on DeepSpeak v1 [1].

top), where Wav2Lip with real audio is left out during training but Wav2Lip with fake audio is not, and (b) evaluation "blind spots", i.e., no performance is reported on FaceSwap and FS-GAN with real audio. We propose a new "method" and "family" based evaluation protocol for FakeAVCeleb (similar to above), see Fig. 3 bottom. Again, the "method" includes one type of video manipulation regardless of the audio manipulation, e.g., Wav2Lip real and fake audio form one method whereas FaceSwap and FaceSwap+Wav2Lip form another method. The Lip Synthesis Family Split includes every modification with Wav2Lip (including the combinations with other methods). The Face Animation Family Split consists of all samples that include either FaceSwap or FS-GAN. Since the modifications FaceSwap+Wav2Lip and FS-GAN+Wav2Lip include both techniques, they are part of both family splits (i.e., never seed during training). Our improved protocol results in a stricter and more challenging evaluation setting.

4. Experimental Results

Sampling Strategies. For the sampling strategies proposed before, the hyperparameters set are $N = 16$ and $M = 5$, following [13]. The most resilient approach to the shortcut is subsampling with temporal jittering, as shown in Table 1. Further results in the section employ this sampling strategy. Regarding DeepSpeak v1, by removing the first $300ms$, the performance between trimmed and untrimmed videos drops by ~ 5 AUC points, but it recovers when subsampling is applied during training. The same holds for FakeAVCeleb, where the impact of this sampling strategy is even more visible. In this case, there is a difference in the AUC of ~ 10 for the binary case and ~ 16 for the multiclass. This difference is barely noticeable when we add subsampling with jittering during training.

Cross-manipulation results on FakeAVCeleb. Fig. 4a shows results on FakeAVCeleb [9]. All models perform well on Wav2Lip, the easiest split to generalize to. Supervised models also do well on FS-GAN, revealing strong performance on its realAudio subset. In contrast, FaceSwap realAudio is the hardest, with unimodal models outperforming multimodal ones, likely due to the latter's focus on lip-sync tasks. In the Family Splits (Fig. 4b), models perform well on Lip Synthesis; SIMBA generalizes perfectly. AVAD favors fake audio, suggesting audio reliance. Unimodal models benefit from combined manipulations (e.g., Wav2Lip + face animation), but drop slightly on Wav2Lip alone. This pattern repeats in the Face Animation Split, where FaceSwap remains the most difficult to generalize to. Overall, generalizing from lip synthesis to face animation is harder than the reverse. To sum up, our method and family

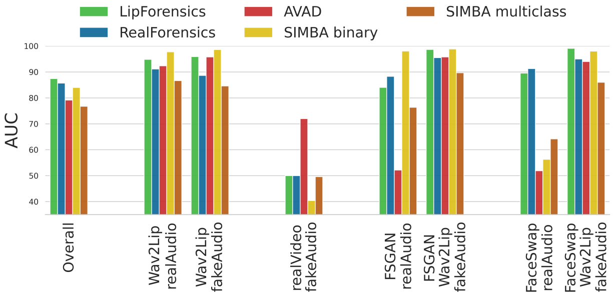


Figure 6: Cross-dataset evaluation across all manipulations, as AUC, from DeepSpeak v1 [1] to FakeAVCeleb [9].

splits show that *FakeAVCeleb is less saturated than previous literature suggests [7, 8, 13]*, re-opening the possibility to use this dataset for cross-manipulation generalization.

Cross-manipulation results on DeepSpeak v1 Method split results on DeepSpeak v1 [1] (Fig. 5a) show most supervised models achieve $> 90\%$ AUC on lip synthesis and perform well on FaceFusion. FaceFusion GAN is the hardest to generalize to, FaceFusion Live is the easiest. *AVAD [6] only performs well when the audio is fake.* Family split results (Fig. 5b) show lower scores on Lip Synthesis compared to FakeAVCeleb, suggesting FaceSwap and FS-GAN are closer to Wav2Lip than newer FaceFusion manipulations. AVAD excels in fake audio due to its alignment focus. SIMBA outperforms unimodal models in the Lip Synthesis family. On Face Animation, RealForensics [7] outperforms SIMBA. As before, FaceFusion GAN is hardest, Live is easiest. Overall, DeepSpeak v1's manipulations are more challenging than those in FakeAVCeleb.

Cross-dataset generalization Lastly, we present the cross-dataset generalization results (Fig. 6). Models were trained on all manipulations from the DeepSpeak v1 dataset and tested on FakeAVCeleb. AVAD [6] performs the worst when the audio is real, but recovers in the cases with fake audio. SIMBA struggles on realVideo-fakeAudio splits, while it shows strong performances on Wav2Lip. Overall, unimodal models have slightly better generalization performances than SIMBA and AVAD. We see that *jointly generalizing to a new manipulation and a new dataset is challenging yet not impossible for SoTA methods*.

5. Conclusion

We contribute to the multimodal DeepFake detection by diagnosing the issues of the prior benchmarks, presenting a baseline method SIMBA, and bringing the recent and challenging dataset DeepSpeak v1 into the spotlight. SIMBA stands out by its simplicity, yet it is competitive with SoTA architectures. *We will make the code publicly available.* We analyze the promising recent dataset, DeepSpeak v1, where we benchmark SoTA models using our new method- and family-split evaluation protocols, revealing that DeepSpeak v1 is challenging for cross-manipulation generalization. On FakeAVCeleb, we expose flaws in the established evaluation protocol and suggest the use of our method and family splits to offer more realistic generalization scenarios. Finally, our simple jittering augmentation scheme is effective at countering the leading silence shortcuts.

References

[1] S. Barrington, M. Bohacek, and H. Farid. Deepspeak dataset v1.0. *CoRR*, abs/2408.05366, 2024. 1, 2, 3, 4

[2] D. Boldisor, S. Smeu, D. Oneata, and E. Oneata. Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning. *To appear in CVPR*, 2025. 1, 2

[3] H. Cheng, Y. Guo, T. Wang, Q. Li, X. Chang, and L. Nie. Voice-face homogeneity tells deepfake. *ACM Trans. Multim. Comput. Commun. Appl.*, 20(3):76:1–76:22, 2024. 1

[4] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian. Not made for each other- audio-visual dissonance-based deepfake detection and localization. In *ACM Multimedia*, pages 439–447. ACM, 2020. 1

[5] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton-Ferrer. The deepfake detection challenge dataset. *CoRR*, abs/2006.07397, 2020. 1

[6] C. Feng, Z. Chen, and A. Owens. Self-supervised video forensics by audio-visual anomaly detection. In *CVPR*, pages 10491–10503. IEEE, 2023. 1, 4

[7] A. Haliassos, R. Mira, S. Petridis, and M. Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *CVPR*, pages 14930–14942. IEEE, 2022. 1, 4

[8] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *CVPR*, pages 5039–5049. Computer Vision Foundation / IEEE, 2021. 1, 4

[9] H. Khalid, S. Tariq, M. Kim, and S. S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. In *NeurIPS Datasets and Benchmarks*, 2021. 1, 2, 3, 4

[10] P. Kwon, J. You, G. Nam, S. Park, and G. Chae. Kodf: A large-scale korean deepfake detection dataset. In *ICCV*, pages 10724–10733. IEEE, 2021. 1

[11] W. Liu, T. She, J. Liu, B. Li, and ... Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes. In *NeurIPS*, 2024. 1

[12] D. Niizumi, D. Takeuchi, Y. Ohishi, and ... BYOL for audio: Self-supervised learning for general-purpose audio representation. In *IJCNN*, pages 1–8. IEEE, 2021. 2

[13] T. Oorloff, S. Koppisetti, N. Bonettini, D. Solanki, B. Colman, Y. Yacoob, A. Shahriyari, and G. Bharaj. AVFF: audio-visual feature fusion for video deepfake detection. In *CVPR*, pages 27092–27102. IEEE, 2024. 1, 3, 4

[14] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459. Computer Vision Foundation / IEEE Computer Society, 2018. 1

[15] W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, and ... Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Trans. Inf. Forensics Secur.*, 18:2015–2029, 2023. 1