The *DeepSpeak* Dataset

Sarah Barrington[1], Maty Bohacek[2], Hany Farid[1]
[1]University of California, Berkeley, [2]Stanford University

Abstract

Deepfakes represent a growing concern across domains such as disinformation, fraud, and non-consensual media. In particular, the rise of video conference and identity-driven attacks in high-stakes scenarios–such as impostor hiring–demands new forensic resources. Despite significant efforts to develop robust detection classifiers to distinguish the real from the fake, commonly used training datasets remain inadequate: relying on low-quality and outdated deepfake generators, consisting of content scraped from online repositories without participant consent, lacking in multimodal coverage, and rarely employing identity-matching protocols to ensure realistic fakes. To overcome these limitations, we present the DeepSpeak dataset, a diverse and multimodal dataset comprising over 100 hours of authentic and deepfake audio-visual content, specifically focused on the challenging and diverse "talking heads" context. We contribute: i) more than 50 hours of real, self-recorded data collected from 500 diverse and consenting participants, ii) more than 50 hours of state-of-the-art audio and visual deepfakes generated using 14 video synthesis engines and three voice cloning engines, and iii) an embedding-based, identity-matching approach to ensure the creation of convincing, high-quality identity face swaps that realistically simulate adversarial deepfake attacks. We also perform large-scale evaluations of state-of-the-art deepfake detectors and show that, without retraining, these detectors fail to generalize to this DeepSpeak dataset, highlighting the importance of a large and diverse dataset containing deepfakes from the latest generative-AI tools.

https://github.com/hfaridlab/deepspeak

1. Introduction

Today, generative-AI is capable of creating hyper-realistic images [36], voices [4], and videos [19] of people talking or doing just about anything. These technologies hold the promise to both revolutionize many industries while also amplifying the spread and belief in dangerous lies and conspiracies [10, 49], interfering with elections [16, 45], supercharging small- and large-scale fraud [5], and – seemingly unable to escape its roots – continue to be used in the creation of non-consensual intimate imagery (NCII) [13, 50].

Scalable, generalizable, and accurate detection of deepfakes has, therefore, become a pressing problem with deep social, political, and economic implications. At the same time, the nascent digital-forensics community has struggled with the lack of large-scale, high-quality, up-to-date, and ethically collected datasets for training and evaluation.

In this work, we introduce an audio and video dataset designed to aid the digital-forensic, computer-vision, and broader AI-safety communities. This dataset consists of 100 hours of real and deepfake video of people talking and gesturing. The real videos were self-recorded with consent from the participants using their own hardware, ensuring a wide range of recording environments, hardware variability and identities, a crucial component for the development of robust detectors. These deepfakes consist of avatar deepfakes (from three generators), face-swap deepfakes (with multiple variants from three generators), lip-sync deepfakes (from four generators), and audio deepfakes (from three generators) spliced into a subset of the lip-sync deepfakes. Table 1 presents a comparison between our dataset and recent datasets released over the past seven years.

We focus exclusively on "talking heads" in which one person is talking or gesturing in a typical video conferencing setup. We focus on this context because of the increasing prevalence of distinct harms that have emerged from both offline and real-time deepfake identity impersonations (including impostor hiring, fraud, and disinformation). This scenario presents unique challenges concerning identity verification and liveness detection, as compared to the more common analysis of text- or image-to-video content. The "talking heads" scenario shifts the question from only a 'real v. fake' or 'identity' question to both one of realism *and* identity. Our effort provides a dataset to simultaneously address both of these questions. Existing datasets in this domain do not reflect the breadth of ours, nor additionally, the current quality and diversity of deepfake generators, while bearing ethical, practical, and legal shortcomings (see Table 1). Specifically, existing datasets largely comprise low-quality, outdated deepfake generators, where the underlying data was scraped without participant consent. Moreover,

Figure 1. An overview of the *DeepSpeak* Dataset sourced from a diverse selection of consenting participants using a custom-built data collection methodology. The dataset also comprises deepfakes generated from 14 video and three audio deepfake methods using facial identity matching to improve the realism of the generated deepfakes.

these datasets do not include all types of deepfake generators and attack settings. A more comprehensive review of other datasets is included in Appendix B.

To remedy these shortcomings, our work makes the following contributions:

- **Documentation and Release of DeepSpeak.** We introduce a methodology for the large-scale collection of real video recordings, self-submitted by a diverse selection of consenting participants (Section 2), along with the procedures used to generate corresponding deepfake video and audio (Sections 4 and 5).
- **Data Collection Tool.** We provide a codebase for a web-based application designed to facilitate participant-led remote data collection. When used in conjunction with our collection survey, the collected data is phonetically rich and diverse in terms of speech content, video durations, gestures, and includes both scripted and unscripted segments.
- **Method for Identity Matching.** We devise a method for matching participants based on their visual features to create more convincing face-swap deepfakes, consistent with real-world deepfake attacks (Section 3).
- **Large-scale Benchmarking and Generalization Study.** We perform large-scale evaluations of state-of-the-art deepfake detectors across audio, visual and multimodal detectors and show that they fail to accurately distinguish between real and fake audio and video when trained on other datasets (Table 1)(Section 6). These evaluations highlight the importance of a large and diverse dataset containing deepfakes from the latest generative-AI tools.

2. Data Collection

The data collection was performed in four steps. Data collection for *DeepSpeak* was determined to qualify for exempt status by UC Berkeley Office for Protection of Human Subjects (OPHS).

Participant Recruitment. Participants were crowdsourced through the Prolific research recruitment platform. Participants were asked to give their consent for including their recordings, without any other identifying information, in a public dataset. Details of the consent statement can be found in Appendix O. A total of 500 participants were selected from a stratified sample ensuring equal distribution of gender, and with all participants reported as being native English speakers and U.S. residents, with demographics as follows (some participants identified with more than one race/ethnicity):

Name	Release Year	Unique Identities	Original Footage	Consent	Faceswap	Lipsync	Avatar	Fake Audio	Conversational Webcam	Identity Matching	Deepfake Footage
FaceForensics	2018	NA	1,004	N	?	-	-	-	-	-	2,008
FaceForensics++	2019	NA	1,000	Partial	✓	-	-	-	-	-	4,000
DFDC	2020	3,426	23,654	Y	✓	-	-	-	Partial	✓	104,500
DFD	2019	28	363	Y	✓	-	-	-	✓	-	3,068
Celeb-DF	2020	59	590	N	✓	-	-	-	-	-	5,049
AVDeepfake1M	2024	2,068	286,721	N	-	✓	-	✓	-	-	860,039
FakeAVCeleb	2021	600	570	N	✓	✓	-	✓	-	-	25,000
Deepfake-Eval-2024	2025	NA	–	N	✓	-	-	✓	Partial	-	–
LAV-DF	2022	153	36,431	N	-	✓	-	✓	-	-	99,873
DF40	2024	NA	NA	N	✓	✓	✓	-	-	-	100,000+
NVFAIR	2025	161	NA	Y	-	✓	-	-	✓	-	650,000
Polyglotfake	2024	NA	766	N	-	✓	-	✓	-	-	14,472
Illusion	2025	NA	139,740	N	✓	-	-	✓	-	✓	1,232,246
DF-Platter	2023	454	764	N	✓	-	-	-	-	-	132,496
DeepSpeak (Ours)	2025	500	16,043	Y	✓	✓	✓	✓	✓	✓	14,005

Table 1. A comparison of forensic-themed public datasets. Although not the most informative metric, we report original and deepfake footage as number of videos for consistency with previous published datasets (NA: not available). "Fake Audio" refers to speech synthesized by AI-enabled voice cloning.

- **Age**: Range = 18-75 years, Mean = 38 years; standard deviation = 11.5 years
- **Gender**: 256 male, 235 female, 7 non-binary, 2 not provided
- **Race/Ethnicity**: 362 White/Caucasian, 87 Black/African American, 45 Asian, 14 American Indian/Alaska Native, 2 Native Hawaiian/Other Pacific Islander, 15 other, 1 prefer not to say.

Survey. The data collection survey was designed to capture both speech and visual actions. For speech, it included phonetically rich audio data spanning varied audio durations with both scripted vs. conversational-style responses. Each participant was instructed to record themselves responding to between 32 and 35 separate prompts. Participants were paid $7 for their time. The first two prompts were used for voice-clone training data (see Section 4). The remaining prompts were divided into four categories: (1) 10 standardized scripted responses in which each participant read the same prompt; (2) 10 randomized scripted responses in which participants read a randomized prompt; (3) 10 unscripted responses in which participants responded to questions; and (4) between 5-8 actions in which participants performed simple actions. Scripted responses were generated using transcripts of the TIMIT dataset, a linguistics research dataset consisting of utterances from 462 real female and male American-English speakers. See Appendix P for the full list of prompts and scripts used.

Data Collection tool. Both audio and video were recorded using a custom-built Google Chrome web application. The JavaScript and Python repository for this web application is available at https://github.com/hfaridlab/deepspeak. Details of the encoding and data preprocessing associated with the tool can be found in the Appendix C.

Validation. Participants were given written and visual instructions to allow them to practice recording themselves and test their hardware. Participants were asked to adhere to a series of recording conditions intended to improve consistency within the overall dataset. We manually removed any invalid responses from the final dataset that did not meet these requirements. The details of this can be found in the Appendix C.

3. Identity Matching

During manual inspection of the collected data, we observed that, albeit diverse in age, gender, and ethnicity, our collected data contains many individuals with similar facial and vocal features. In order to exploit this feature of the dataset and create more compelling deepfakes, each identity in the dataset was paired with another, perceptually similar one. The code for producing this visual matching, as well as the resulting visual pairs is open-sourced at https://github.com/hfaridlab/deepspeak.

Visual Matching. Each identity is first represented by the average CLIP embedding[1] [39] extracted from five random video frames (filtered for low-quality frames, see Section 5.2). Shown in Figure 2 is a t-SNE visualization of a subset of these embeddings. Comparing this representation against the self-reported demographic information reveals that these CLIP embeddings cluster based on gender, ethnicity and facial similarity.

[1] https://github.com/OpenAI/CLIP

Figure 2. A t-SNE visualization of CLIP embeddings from real participant's videos. The four highlighted pairs correspond to identities with maximal similarity as measured by the cosine distance between CLIP embeddings. Perceptually similar identities cluster in this t-SNE representation. The red/blue color coding corresponds to people who identify as female/male, which also clusters in this t-SNE representation.

For each identity, a unique matched identity is assigned using the agglomerative clustering algorithm with cosine distance and cluster size constraint from the scikit-learn library[2]. Additional examples of visual pairs are shown in Appendix I. This approach was adopted instead of a more traditional biometric matching like ArcFace [12] because we observed, during manual review, better qualitative matching for women and people of color. We also found that our CLIP-based matching outperforms ArcFace in terms of the Frechet inception distance (FID) between the matched face pairs by 4% (214 vs 222), and the LPIPS distance between the matched face pairs by 6% (0.61 vs 0.65). For both FID and LPIPS, a smaller value corresponds to higher perceptual similarity.

With this identity matching, averaged across all videos, DeepSpeak achieves an average FID of 238 and LPIPS of 0.46. Compared to baselines datasets, this is 36% better than DF40 (325 FID and 0.68 LPIPS), 33% better than FaceForensics (317 and 0.74), and 71% better than DFDC (408 and 0.70).

4. Audio Generation

Participants were first asked to record themselves reading 10 consecutive phonetically-rich sentences, sourced from List 1 of the standard Harvard Sentences [42], a collection of sentences representing best practice for standardized evaluation of speech processing and audio quality in controlled settings. Participants were then asked to repeat the standard elicitation paragraph from the Speech Accent Archive, a phonetically comprehensive passage comprising a breadth of vowels and consonants [54]. These two scripted responses were used for the purpose of voice cloning, and had an average length of 30 seconds.

Using each participant's cloned voice, a synthetic audio was created in their voice saying the same thing as in the original audio/video. For the unscripted responses, the original audio was transcribed using OpenAI's Whisper, and for the scripted responses, we assumed that the participant correctly read the script. These text transcriptions were then provided to each voice cloning generators' API to generate matching synthetic voices.

Voice clones were generated using three commercial cloning and Text-to-Speech (TTS) services: ElevenLabs, PlayAI and Speechify. The details of API end points used, alongside parameters, can be found in Appendix E.

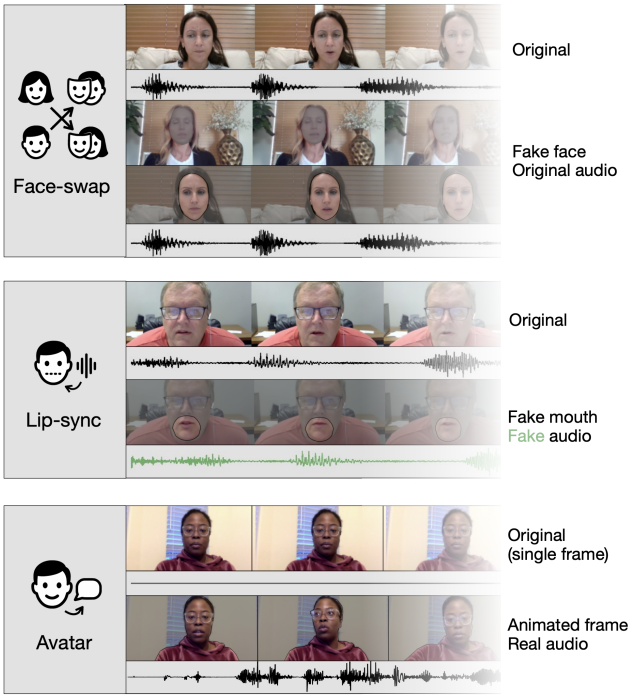

Figure 3. The *DeepSpeak* dataset consists of face-swap, lip-sync, and avatar deepfakes.

5. Video Generation

We generated three types of video deepfakes: face swap, lip sync, and avatar, each of which is described next. The resulting dataset is randomly split into 80/20 training/testing splits with no overlap in facial or voice identities. A breakdown of the resulting dataset's statistics, including the total file size (GB), file counts (N), and video length (hrs) are included in Appendix D.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

5.1. Generation

Face-Swap. Face-swap deepfakes are created by replacing – eyebrow to chin and cheek to cheek – the original identity in a video with a new identity. We swapped faces of identity pairs identified through the visual matching (see Section 3). This ensured that the swapped identities were perceptually similar to begin with, which made for more compelling deepfakes. This resembles conventional practices of in-the-wild deepfake production, where actors are chosen based on their similarity to the target identity.

An overview of face-swap deepfake generation is shown in Figure 3, row one. To generate a face swap, the video of the original identity and a single frame of the matched identity are provided to the face-swap synthesis engine. The single frame is initially chosen to be the fifth frame in a randomly selected video of the matched identity. We found that if the eyes are closed in the matched face, the resulting face-swap deepfake suffered in quality. As such, we used MediaPipe [33] to extract facial features and ensured that the distance between the top and bottom eyelid landmarks was greater than a specified threshold. If this constraint failed, then the tenth frame was selected for consideration; this process was repeated, skipping five frames each time, until a suitable frame was found. We used seven face-swap methods, as detailed in Appendix D.

Lip-Sync. Whereas the face-swap deepfake replaces an entire face with a new identity, a lip-sync deepfake modifies the mouth region to be consistent with a different audio track. An overview of lip-sync deepfake generation is shown in row two of Figure 3. Given an original video and associated audio, we create two types of lip-sync deepfakes: (1) a lip-sync deepfake with an audio of the same identity extracted from a different video (i.e., the audio and video are now mismatched); and (2) a lip-sync deepfake with an AI-generated voice of the same identity (Section 4) with a transcript taken from a different video. Four methods of lip-sync deepfakes were employed, as further described in Appendix D.

Avatar. An overview of avatar deepfake generation is shown in Figure 3, row three. Avatar deepfakes animate the head and lip movements of a static image to match a target video or audio track. Unlike face-swap and lip-sync deepfakes, which modify an existing video, avatar deepfakes generate movement from a single static image. Avatar deepfakes were created using three methods further described in Appendix D. LivePortrait and HelloMeme take as input a single image of a person to animate with a video (and associated audio) that drives this animation. For these two generators, the avatar deepfakes contain only real audio from the original driving video. Memo takes as input a single image of a person to animate with only an audio that drives this animation. In this case, the audio can be either real or fake.

5.2. Validation

During manual inspection of the generated videos, we identified multiple types of failures, including deepfake engines (1) producing corrupted faces with consistently closed eyes or mouths, (2) generating malformed avatars with distorted facial or upper body structure, (3) failing to apply any changes and yielding back the original video, (4) modifying only parts of the video, (5) producing empty output consisting of with black frames, among others. To prevent failed deepfakes, we designed a suite of input and output detectors to filter undesired features. This filtering code is open-sourced at https://github.com/hfaridlab/deepspeak. The details of this filtering can be found in Appendix F.

6. Experiments

We conducted a series of baseline experiments on *DeepSpeak* for the tasks of audio and video deepfake detection. The code for these experiments, including data pre-processing, is open-sourced at https://github.com/hfaridlab/deepspeak. The experiments were conducted on NVIDIA A100 GPUs over the course of approximately four weeks (see Appendix N for details pertaining compute resources).

6.1. Video Deepfake Detection

Baselines. Both classic- and deep-learning methods for deepfake video detection can be categorized by the scrutinized signal deemed to discriminate the real from the fake, with most performing (1) spatial-domain analysis, (2) frequency-domain analysis, or (3) cross-modal temporal coherence analysis. To capture the breadth of the existing approaches, we evaluate state-of-the-art methods representing these distinct lines of work. The first evaluated architecture is a frequency-based method FreqNet [7]. The second, spatial-domain, architecture is GenConViT [55] (with ED and VAE variants). The third, multi-modal, architecture is LipFD [30] designed to detect misalignments between the visual and vocal stream of lip-sync deepfakes.

For each of these four architectures, we evaluated three model variants: (1) the pretrained model released alongside the respective publication (trained on a different, non-DeepSpeak dataset), (2) the model trained from scratch on DeepSpeak, and (3) the model, starting with the pre-trained weights, fine-tuned on DeepSpeak. A total of 12 models were evaluated.

Experimental Setup. To perform inference, training, and fine-tuning of the included architectures, we used the official code repositories released alongside the respective publications. To make the results comparable despite the differing number of parameters of these architectures, we used default hyperparameters when possible, with a simple search over learning rates (see Appendix K for details).

Each model is evaluated against the testing split of its architecture's original dataset and DeepSpeak. The original dataset refers to the dataset used for the pretrained model in the respective publication: for FreqNet, it is a custom GAN-generated dataset compiled by its authors [7]; for GenConViT ED and VAE, it is Celeb-DF 2 [28]; and for LipFD, it is AVLips [30]. The accuracy on the real and fake class is reported separately, along with the overall F1 score.

Results. Shown in Table 2 are the results of the pretrained models and models trained from scratch on DeepSpeak. All four evaluated architectures follow the same pattern: they perform reasonably well on the testing splits of their original training datasets but fail to generalize to DeepSpeak. The same trend holds when models are trained on DeepSpeak and evaluated on the original dataset. Notably, even on the original testing sets, class bias was evident—for example, GenConViT attained an accuracy of 98.2% on fake but only 56.7% on real, while LipFD showed the opposite pattern, scoring 97.9% on real versus 69.1% on fake.

Also shown in last six columns of Table 2 are the results of the fine-tuned models (labeled Original + DeepSpeak). While some models, such as GenConViT ED and VAE, achieved performance on DeepSpeak comparable to training from scratch (F1 score above 0.9), this came at the cost of a sharp drop in performance on the original testing set, where F1 scores fell below 0.2. LipFD was able to fine-tune on DeepSpeak while maintaining comparable performance on the original testing set (both with F1 scores around 0.7), though it should be noted that the model exhibits a strong bias toward the fake class.

6.2. Audio Deepfake Detection

Baselines. We evaluated the performance of two model architecture types on the DeepSpeak dataset, consistent with recent literature: (i) a foundation model, and (ii) a raw waveform model. Foundation models use a pretrained model to extract embeddings from the input waveform, which are then passed to a classifier. Three state-of-the-art models were selected: TitaNet [3, 25], Wav2Vec-XLSR [2, 37], and LAION-CLAP [37, 56]. For each embedding type, both linear and non-linear classifiers were tested. Raw waveform models operate directly on the audio waveform. Three leading models were chosen: AASIST [23], RawNet2 [23, 48], and RawGAT-ST [23, 47].

For both architectures, we evaluated two versions of each model: (1) a pretrained model trained on a dataset other than DeepSpeak, and (2) a model trained from scratch on DeepSpeak. In the case of foundation models, the foundation model used to extract embeddings remained pretrained, while the downstream classifiers were trained from scratch. In total, 18 models were evaluated. A summary is provided in Table 6.2.

Experimental Setup. Pretrained raw waveform model weights were sourced directly from the AASIST implementation of AASIST, RawNet2, and RawGAT-ST [3]. Default configuration were used for each model, as detailed in the Appendix K. For retraining these models from scratch on DeepSpeak, the same configurations and architectures were maintained, with DeepSpeak training data replacing ASVSpoof. For foundation models, classifiers (both logistic regression and random forest) were trained using balanced datasets with embeddings extracted from the training sets of either ASVSpoof (for Wav2Vec-XLSR and LAION-CLAP) or TIMIT-ElevenLabs (for Titanet). Embeddings from the DeepSpeak test dataset split were used for evaluation. Both linear (logistic regression) and non-linear (random forest) classifiers were tested for each embedding type. No cross-validation or hyperparameter tuning was performed for either the pretrained or from-scratch models.

Each model is evaluated against the testing split of its architecture's original dataset and DeepSpeak. The original dataset corresponds to the one used for pretraining in the respective publications. For AASIST, RawNet2, and RawGAT-ST, this dataset is ASVSpoof (as implemented in [23]), and for TitaNet-based embeddings approaches, this dataset is TIMIT-ElevenLabs [3]. Since prior literature on detection using Wav2Vec-XLSR and LAION-CLAP largely focusses on training-free methods [37], we trained our own benchmarks on ASVSpoof for consistency and because it serves as one of the most comprehensive and widely used benchmarking datasets. Performance metrics are reported for both the original dataset's test set and the DeepSpeak test set. As shown in Table 6.2, accuracies for both real and fake classes are presented separately, along with overall accuracy to account for class imbalance (since fake audio only occurs in lip-sync deepfakes, representing a subset of the full dataset), and the error rate (EER).

Results. When trained and tested on DeepSpeak, raw waveform models perform well, with AASIST achieving 98.8% accuracy - only 0.7 percentage points lower than its original ASVSpoof benchmark. Embedding-based models also show strong, though comparatively lower performance, with the best performing models being those trained on LAION-CLAP embeddings (see Table 6.2).

Pretrained models, however, do not generalize well to DeepSpeak data. AASIST remains the top-performing pretrained model, albeit with substantially lower performance when evaluated out-of-the-box on DeepSpeak data, dropping to an accuracy of 60.1% and 60.2% for real and fake. Pretrained embedding-based models also show substantially lower performance when evaluated on DeepSpeak data, alongside notable class imbalances (see Table 6.2).

These results suggest that feature representations learned

[3] https://github.com/clovaai/aasist

	Original						DeepSpeak						Original + DeepSpeak					
	Original			DeepSpeak			Original			DeepSpeak			Original			DeepSpeak		
Method	Real	Fake	F1	Real	Fake	F1	Real	Fake	F1	Real	Fake	F1	Real	Fake	F1	Real	Fake	F1
FN	97.1	88.3	0.9	65.3	15.4	0.2	34.4	26.6	0.6	77.3	69.9	73.6	50.5	14.1	0.3	74.2	66.1	0.7
GC-ED	57.3	98.2	0.7	88.5	39.1	0.7	2.8	100	0.1	90.5	90.7	0.9	7.9	100	0.2	91.7	78.2	0.9
GC-VAE	56.7	98.2	0.7	88.5	39.2	0.7	4.5	100	0.1	91.1	96.4	0.9	9.0	99.7	0.2	93.0	89.6	0.9
LFD	97.9	69.1	0.8	98.8	3.5	0.1	7.30	88.7	0.7	71.8	77.1	0.8	2.8	97.8	0.7	28.2	96.6	0.7

Table 2. **Video deepfake detection accuracies (%)** of four state-of-the-art architectures: FreqNet (FN), GenConViT ED (GC-ED), GenConViT VAE (GC-VAE), and LipFD (LFD). The heading in the first row corresponds to the dataset on which each model was trained, and the heading in the second row corresponds to the dataset on which each model is evaluated against.

		Original						DeepSpeak					
		Original			DeepSpeak			Original			DeepSpeak		
Model	**Clf**	Real	Fake	F1	Real	Fake	F1	Real	Fake	F1	Real	Fake	F1
Titanet (FM)	LR	99.4	100.0	1.00	**10.0**	**97.4**	**0.18**	61.6	98.8	0.75	91.3	89.1	0.95
	RF	99.8	100.0	1.00	**54.2**	**64.3**	**0.68**	74.8	83.1	0.71	96.3	79.3	0.97
Wav2Vec2-xlsr (FM)	LR	79.7	82.7	0.48	**1.3**	**97.0**	**0.03**	7.6	83.8	0.06	76.8	65.6	0.84
	RF	98.1	88.5	0.66	**19.3**	**95.4**	**0.32**	93.6	36.9	0.25	97.4	78.0	0.97
LAION-CLAP (FM)	LR	92.9	92.0	0.71	**33.8**	**76.6**	**0.49**	90.3	53.3	0.30	93.7	91.9	0.96
	RF	93.1	90.5	0.67	**65.3**	**68.3**	**0.77**	93.9	56.7	0.33	95.8	89.6	0.97
AASIST (RW)	-	99.5	99.5	0.98	**60.1**	**60.2**	**0.72**	73.1	73.1	0.36	98.8	98.8	0.99
RawNet2 (RW)	-	99.0	99.0	0.95	**54.4**	**54.4**	**0.67**	69.4	69.3	0.32	94.1	94.3	0.96
RawGAT-ST (RW)	-	99.1	99.1	0.96	**57.7**	**57.7**	**0.70**	75.0	75.0	0.38	96.8	96.8	0.98

Table 3. **Audio deepfake detection accuracies (%)** of nine state-of-the-art models using two separate architectures (FM = Foundation Model, RW = Raw Waveform). For each training/testing combination, we report the real class accuracy, fake class accuracy, and F1 score. The Clf column indicates the type of classifier used for embedding-based models, either logistic regression (LR) or random forest (RF).

directly from raw waveform inputs may be more resilient to domain shift in DeepSpeak data than those extracted from foundation embeddings-based models. In all cases, pre-trained models are insufficient for accurately distinguish real from fake audio.

This pattern for audio detection models is similar to video detection models: (1) these models struggle with out-of-domain data; but (2) these models can improve with appropriate training.

6.3. Combined Audio-Visual Detection

While most deepfake detection methods treat audio and video independently, several joint approaches have also been proposed. Although multimodal detection is not the primary focus of this work, we confirm that the limited generalization observed in unimodal detection methods extends to a leading combined method exploiting audio-video mismatch [6]. This technique classifies a video as real/fake based on a normalized Levenshtein distance between an audio transcription and a video transcription (based on an automatic lip reading). At a threshold of 0.55 on this distance (where are distance of 0.0 corresponds to a perfect match between the audio and video, and a distance of 1.0 corresponds to a maximal mismatch), the accuracy across 2963 real and 2953 AI-generated videos is 65.5% and 67.9%. Accuracy across different types of AI-generated videos (face-swap, lip-sync, and avatar) ranged from a low of 43.4% to a high of 99.0%.

7. Closing Thoughts

Discussion. Year after year, we see a dramatic rise in the number of deepfake generators and the quality of the fake audio and video. Given the pace at which deepfake technology is progressing, it is critical that evaluation datasets keep up with the latest technologies. This is made apparent by our evaluation of recent state-of-the-art deepfake detectors that struggle to generalize to the latest deepfake generators. To this end, our DeepSpeak dataset is partitioned into two parts (v1 and v2), containing a snapshot of the state of the art in deepfake generation in 2024 and 2025, respectively. We plan to release one to two new datasets each year to keep pace with these new threats.

Limitations. *DeepSpeak* captures the state of the art in deepfake generation at the time of publication, making it well-suited for developing and evaluating detection methods for current and emerging deepfake engines. However, as

generative AI evolves rapidly, it is essential to recognize the dataset's limitations and the potential for future expansion.

Due to the lack of high-quality open-source deepfake engines for non-English languages, DeepSpeak currently includes participants speaking English. As high-quality multilingual engines become available, we will need to expand DeepSpeak to include additional languages.

Currently, open-source deepfake engines operate on the video level, which is reflected in DeepSpeak—every video in the dataset is either entirely real or entirely fake. Once targeted manipulation (i.e., changing only some words in the video) improves, we will include them in future versions.

Lastly, to date, all of the DeepSpeak video generators are based on open-source models and not on commercially available models. As we have done with commercial audio generators, we will seek to establish relationships with commercial video generators to allow for large-scale video generation of commercial offerings.

Ethical Considerations: Too many other datasets in media forensics and computer vision have adopted a "scrape and distribute, ask questions later" approach. We take issue with this both from the perspective of participant consent and intellectual property.

While we don't object to the development of deepfake generators, we will not knowingly license *DeepSpeak* for this purpose. Our rationale here is that the harms that are coming from deepfakes are not insignificant and we simply don't want to be contributing to a plethora of online harms.

Conclusion. Our motivation for creating this dataset is to support the media-forensics research community and the development and refinement of techniques to detect deepfake audio, image, and video. The world of generative AI and media forensics is fast moving. It is, therefore, important that shared datasets be regularly updated to keep up with the latest trends. To this end, we expect to release updates to this dataset once to twice a year. To help serve the community better, we welcome feedback, comments, requests for future releases of this dataset at https://github.com/hfaridlab/deepspeak.

Acknowledgments

We are grateful to David Chan for his many insightful comments and suggestions that significantly improved the quality of this paper. This work was supported by Google/YouTube and the University of California Noyce Initiative. We are grateful to ElevenLabs (https://elevenlabs.io) and PlayAI (https://play.ai/) for granting us API access for voice generation.

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8717–8727, 2018. 17

[2] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech*, pages 2278–2282, 2022. 6

[3] Sarah Barrington, Romit Barua, Gautham Koorma, and Hany Farid. Single and multi-speaker cloned voice detection: From perceptual to learned features. In *IEEE International Workshop on Information Forensics and Security*, pages 1–6. IEEE, 2023. 6

[4] Sarah Barrington, Emily A Cooper, and Hany Farid. People are poorly equipped to detect AI-powered voice clones. *Scientific Reports*, 15(1):11004, 2025. 1

[5] Jon Bateman. *Deepfakes and synthetic media in the financial system: Assessing threat scenarios*. Carnegie Endowment for International Peace., 2022. 1

[6] Matyas Bohacek and Hany Farid. Lost in translation: Lip-sync deepfake detection from audio-video mismatch. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4315–4323, 2024. 7

[7] Runyuan Cai, Yue Ding, and Hongtao Lu. FreqNet: A frequency-domain image super-resolution network with discrete cosine transform. 2021. 5, 6

[8] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2003–2011, 2020. 16

[9] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. VideoRetalking: Audio-based lip synchronization for talking head video editing in the wild. In *SIGGRAPH Asia*, pages 1–9, 2022. 17

[10] Bobby Chesney and Danielle Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107:1753, 2019. 1

[11] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep speaker recognition. arXiv:1806.05622, 2018. 17

[12] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 4

[13] Michelle L Ding and Harini Suresh. The malicious technical ecosystem: Exposing limitations in technical governance of ai-generated non-consensual intimate images of adults. arXiv:2504.17663, 2025. 1

[14] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (DFDC) dataset. arXiv:2006.07397, 2020. 14

[15] Nick Dufour and Andrew Gully. Contributing data to deepfake detection research. Google AI Blog, 2019.

[16] Emilio Ferrara. Charting the landscape of nefarious uses of generative artificial intelligence for online election interference. arXiv:2406.01862, 2024. 1

[17] J. H. Frank and L. Schönherr. WaveFake: A Data Set to Facilitate Audio DeepFake Detection. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, pages 1–18, 2021. 14

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. DARPA TIMIT acoustic phonetic continuous speech corpus, 1993. 15

[19] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1):e2110013119, 2022. 1

[20] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. LivePortrait: Efficient portrait animation with stitching and retargeting control. 2024. 17

[21] Keith Ito and Linda Johnson. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017. 14

[22] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 4485–4495, Red Hook, NY, USA, 2018. Curran Associates Inc. 14

[23] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *IEEE International Conference on Acoustics, Ppeech and Signal Processing*, pages 6367–6371, 2022. 6, 30

[24] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 14

[25] Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg. TitaNet: Neural model for speaker representation with 1D depth-wise separable convolutions and global context. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8102–8106. IEEE, 2022. 6

[26] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. Towards automatic face-to-face translation. In *27th ACM International Conference on Multimedia*, pages 1428–1436, 2019. 17

[27] Chunyu Li, Chao Zhang, Weikai Xu, Jinghui Xie, Weiguo Feng, Bingyue Peng, and Weiwei Xing. LatentSync: Audio conditioned latent diffusion models for lip sync. 2024. 17

[28] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020. 6, 14

[29] Mingcong Liu, Qiang Li, Zekui Qin, Guoxin Zhang, Pengfei Wan, and Wen Zheng. BlendGAN: Implicitly GAN blending for arbitrary stylized face generation. *Advances in Neural Information Processing Systems*, 34:29710–29722, 2021. 17

[30] Weifeng Liu, Tianyi She, Jiawei Liu, Boheng Li, Dongyu Yao, and Run Wang. Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes. *Advances in Neural Information Processing Systems*, 37:91131–91155, 2024. 5, 6

[31] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:2507–2522, 2023. 14

[32] Steven R Livingstone and Frank A Russo. The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. 17

[33] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. MediaPipe: A framework for building perception pipelines. arXiv:1906.08172, 2019. 5, 18

[34] Soumik Mukhopadhyay, Saksham Suri, Ravi Teja Gadde, and Abhinav Shrivastava. Diff2Lip: Audio conditioned diffusion models for lip-synchronization. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5292–5302, 2024. 17

[35] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: A large-scale speaker identification dataset. arXiv:1706.08612, 2017. 17

[36] Sophie J. Nightingale and Hany Farid. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8): e2120481119, 2022. 1

[37] Alessandro Pianese, Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Training-free deepfake voice recognition by leveraging large-scale pre-trained models. In *ACM Workshop on Information Hiding and Multimedia Security*, page 289–294, New York, NY, USA, 2024. 6

[38] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *28th ACM International Conference on Multimedia*, pages 484–492, 2020. 17

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3

[40] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 17

[41] Xingyu Ren, Alexandros Lattas, Baris Gecer, Jiankang Deng, Chao Ma, and Xiaokang Yang. Facial geometric detail recovery via implicit representation. In *IEEE 17th International Conference on Automatic Face and Gesture Recognition*, 2023. 15, 16

[42] Ernst H Rothauser. Ieee recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3):225–246, 1969. 4

[43] Henry Ruhs. FaceFusion. https://github.com/facefusion/facefusion, 2024. 15

[44] Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. JSUT corpus: Free large-scale Japanese speech corpus for end-to-end speech synthesis, 2017. arXiv preprint. 14

[45] Sam Stockwell, Megan Hughes, Phil Swatton, Albert Zhang, Jonathan Hall KC, and Kieran. AI-enabled influence operations: Safeguarding future elections. Technical report, Centre for Emerging Technology and Security (CETaS), The Alan Turing Institute, 2024. 1

[46] Kim Sung-Bin, Lee Chae-Yeon, Gihun Son, Oh Hyun-Bin, Janghoon Ju, Suekyeong Nam, and Tae-Hyun Oh. MultiTalk: Enhancing 3D talking head generation across languages with multilingual video dataset. arXiv:2406.14272, 2024. 17

[47] Hemlata Tak, Jee-Weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. arXiv:2107.12710, 2021. 6

[48] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with RawNet2. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6369–6373, 2021. 6

[49] Cristian Vaccari and Andrew Chadwick. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social media+ society*, 6(1):2056305120903408, 2020. 1

[50] Marco Viola and Cristina Voto. Designed to abuse? deepfakes and the non-consensual diffusion of intimate images. *Synthese*, 201(1):30, 2023. 1

[51] Haofan Wang. INSwapper: Face swapping model based on insightface. https://github.com/haofanwang/inswapper, 2023. 16

[52] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer, 2020. 17

[53] Zhouxia Wang, Jiawei Zhang, Tianshui Chen, Wenping Wang, and Ping Luo. RestoreFormer++: Towards real-world blind face restoration from undegraded key-value pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 16

[54] Steven Weinberger. Speech accent archive, 2015. Retrieved from the Speech Accent Archive. 4

[55] Deressa Wodajo, Solomon Atnafu, and Zahid Akhtar. Deepfake video detection using generative convolutional vision transformer. 2023. 5

[56] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE, 2023. 6

[57] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. VFHQ: A high-quality dataset and benchmark for video face super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022. 17

[58] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. DF40: Toward next-generation deepfake detection. arXiv:2406.13495, 2024. 14

[59] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen. LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–8. IEEE, 2019. 17

[60] Shengkai Zhang, Nianhong Jiao, Tian Li, Chaojie Yang, Chenhui Xue, Boya Niu, and Jun Gao. HelloMeme: Integrating spatial knitting attentions to embed high-level and fidelity-rich conditions in diffusion models. 2024. 17

[61] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 17

[62] Longtao Zheng, Yifan Zhang, Hanzhong Guo, Jiachun Pan, Zhenxiong Tan, Jiahao Lu, Chuanxin Tang, Bo An, and Shuicheng Yan. MEMO: Memory-guided diffusion for expressive talking video generation. 2024. 17

[63] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022. 16

[64] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *NeurIPS*, 2022. 16, 17

[65] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *European Conference on Computer Vision*, pages 650–667. Springer, 2022. 17