THE DEEPSPEAK DATASET

Hugging Face | Paper

Sarah Barrington[1], **Maty Bohacek**[2], **Hany Farid**[1]
University of California, Berkeley[1]; Stanford University[2]
{sbarrington}{hfarid}@berkeley.edu, maty@stanford.edu

CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivations

Deepfakes are an increasing threat across domains including disinformation, fraud, and identity-based attacks such as impostor hiring in video conferencing settings. **Existing deepfake detection datasets are often limited** by outdated generation methods, non-consensual data collection, and other challenges (Fig.1, Table 1). Consequently, **state-of-the-art detection algorithms trained on these datasets often fail to generalize** to newer and more realistic deepfake generators. We introduce DeepSpeak: a **diverse multimodal dataset containing over 100 hours of authentic and deepfake audiovisual "talking head" content,** designed to serve as a challenging benchmark for the development and evaluation of robust multimodal deepfake detectors.

Dataset	Release Year	Fake Video	Fake Audio	Consent	Identity Matched
DFDC	2020	✓	✗	✓	✓
FakeAVCeleb	2021	✓	✓	✗	✗
DF40	2024	✓	✗	✗	✗
...					
DeepSpeak	2025	✓	✓	✓	✓

Table 1: deepfake datasets summary.

Fig.1: Example content from other deepfake datasets [1, 2].

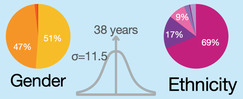
Real Data Collection

1. Participant Recruitment

500 paid crowdworkers
~30 videos each

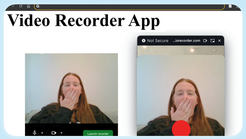
US-based, English first language

Consent statement & additional safeguards

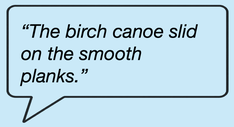


Diverse gender, ethnicities & age

2. Video and Audio Capture



Custom recording tool [open-sourced]



Phonetically rich scripted & unscripted prompts

Multiple gestures

Deepfake Creation

Fig.2: A t-SNE visualization of CLIP embeddings from real participants' videos. Perceptually similar identities cluster in this t-SNE representation.

DeepSpeak comprises **three categories of video deepfakes** (Fig.3), alongside **AI-generated cloned speech** created using commercial voice cloning systems from ElevenLabs, PlayAI, and Speechify.

Face-swap deepfakes were generated by matching visually similar participant identities using CLIP embeddings before applying multiple state-of-the-art synthesis pipelines. Embeddings were found to cluster around visually similar attributes (Fig.2).

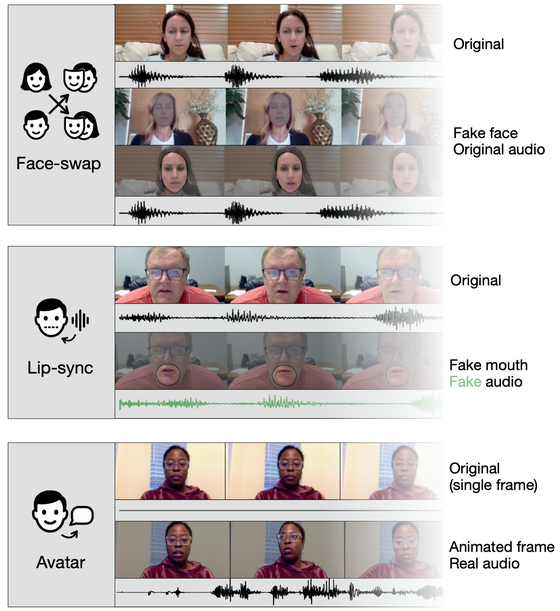


Fig.3: Deepfake creation methods spanning face-swap, lip-sync, avatar, and AI-powered voice cloned audio.

Dataset Summary

Fig.4: Example frames from the DeepSpeak dataset.

- **500** participants
- **>100 hours** real & fake footage
- **>30,000** videos
- **14** video synthesis enines
- **3** voice cloning platforms
- Ongoing releases (v1/2/3 2024/2025/2026)

Experiments

We find that **state-of-the-art deepfake generators fail to generalize to the DeepSpeak dataset**, demonstrating that the dataset provides a new challenge for benchmarking.

Modality	Without Retraining				With Retraining			
	Highest Accuracy		Lowest Accuracy		Highest Accuracy		Lowest Accuracy	
	Real	Fake	Real	Fake	Real	Fake	Real	Fake
Video	98.8	39.2	65.3	3.5	91.1	96.4	71.8	69.9
Audio	65.3	97.4	1.3	54.4	98.8	98.8	76.8	65.6
Combined Audio-Visual	65.5	67.9	*Subtype range: 43.4-99.0*		*N/A*	*N/A*	*N/A*	*N/A*

Table 2: Deepfake detection model accuracies (%) of nine state-of-the-art audio models and four video models.

Experiments were conducted using a range of leading video, audio, and audio-visual (combined) detection models; including FreqNet [3] and GenCon ViT [4] (video), RawNet2 [5] and Titanet embeddings with classifier head [6] (audio), and lip-sync mismatch [7] (combined).

Discussion & Limitations

- Current state-of-the-art deepfake detectors struggle to generalize to newer, more realistic generative models, highlighting the need for continuously updated forensic benchmarks.
- DeepSpeak is designed as an evolving dataset, with planned future releases to track advances in generative AI systems and emerging attack vectors.
- The current release is limited to English-speaking participants due to the lack of high-quality multilingual deepfake generation tools.
- Dataset access is restricted and licensed only for defensive research and reproducibility efforts to reduce misuse risks.

We are grateful for our collaboration with ElevenLabs and PlayAI for voice generation, to Google/YouTube and the University of California Noyce Initiative for funding this work, and to David Chan for his many insightful comments.

References

[1] Lee, Rosemary. (2020). Machine Learning and Notions of the Image; [2] Li, Yuezun, et al. "Celeb-df: A large-scale challenging dataset for deepfake forensics." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020; [3] Runyuan Cai, Yue Ding, and Hongtao Lu. FreqNet: A frequency-domain image super-resolution network with dicrete cosine transform. 2021; [4] Deressa Wodajo, Solomon Atnafu, and Zahid Akhtar. Deepfake video detection using generative convolutional visiontransformer. 2023. 5; [5] 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; [6] 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 [7] Matyas Bohacek and Hany Farid. Lost in translation: Lipsync deepfake detection from audio-video mismatch. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4315–4323, 2024. 7