Temporally Streaming Audio-Visual Synchronization for Real-World Videos - Supplement

Jordan Voas* Wei-Cheng Tseng* Layne Berry Xixi Hu Puyuan Peng
James Stuedemann David Harwath
The University of Texas at Austin
{jvoas, raytseng, layne.berry, hxixi, pyp, jbstuedemann, harwath}@utexas.edu

1. Dataset

1.1. Identifying Talking Heads

The process for identifying talking-head segments in our dataset is illustrated in Figure 1. Following the methodology described in [2], we begin by detecting and cropping faces in the video, generating multiple video clip tracks featuring cropped faces. For each track and its corresponding audio signal, we apply a state-of-the-art lip-syncing model to determine if the audio and video streams are synchronized. Synchronization indicates that the visible individual is the speaker, thus labeling the segment as a talking-head. Conversely, a lack of synchronization suggests the individual is not the speaker, and such tracks are classified as voiceovers. We utilize the S3FD face detector [3] for face detection, which requires a minimum face size of 20 pixels and a minimum track duration of 4 seconds. For lip-syncing analysis, the SyncNet model [2] is employed, with synchrony determined based on a confidence threshold above 0.5 and an absolute offset of less than 5 frames (0.2 seconds).

1.2. Recognizing Audio Events

Figure 2 depicts the process for recognizing audio events in each video clip. The audio stream is segmented into overlapping chunks, each 10 seconds in length, to align with the standard configuration of contemporary audio classifiers. These chunks are then analyzed by an audio classifier to predict the probability of various audio events, with a 5-second overlap between chunks. For audio classification, we use the pre-trained BEATs model [1].

2. Success/Failure Cases

We present examples under three distinct scenarios in the supplementary zip file, alongside descriptions in this section for talking heads, voiceovers, and others. The included video files are those used to evaluate the StreamSync model but have been compressed to reduce upload file sizes.

In the "Talking Heads" scenario, the model relies on speech signals and corresponding lip movements for synchronization, as shown in examples (video no.1-2). However, challenges arise when multiple faces are present in the scene (video no.3-4), requiring speaker identification by the model, or when periodic head movements potentially distract the model (video no.5).

Conversely, the "Voiceover" scenario includes several false negatives, often due to speakers being in challenging visibility conditions, such as facing away from the camera (video no.6) or having obscured facial features (video no.7). These instances highlight the limitations of current models. Moreover, failure cases with actual voiceovers occur when the model misinterprets synchronization cues like the presence of a single clear talking head, resulting in inaccurate predictions (video no.8-9). This observation is consistent with the findings in the "Talking Heads" scenario failure cases.

Finally, in the "Others" scenario, the model effectively utilizes significant, albeit sparse, video events for accurate predictions, such as scene transitions (video no.10), golf ball strikes (video no.11), or tennis ball hits (video no.12). However, it exhibits limitations in processing smoother video events, like musical performances with associated finger movements (video no.13) or athletes sprinting with background applause and chanting (video no.14).

3. Talking Head Bias from Pretrained Models

To compare the resulting bias towards talking heads that RealSync induces we perform the same evaluations on the test set of RealSync and display the performance splits across the different video categories of talking heads, voiceovers, and others in 2. It can be observed that SparseSync, which was pretrained on LRS3 and then finetuned with VGGSound, actually presents a more significant bias towards talking heads than the results seen for StreamSync

*Equal Contribution First Author.

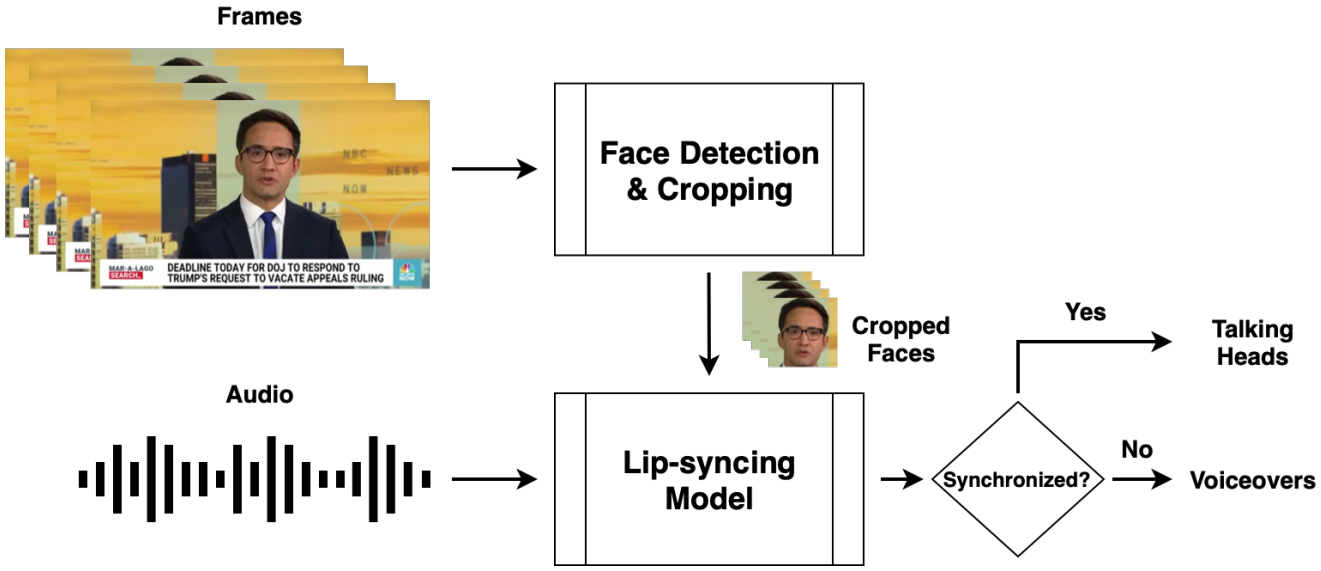

Figure 1. Pipeline overview for identifying talking-head segments in videos.

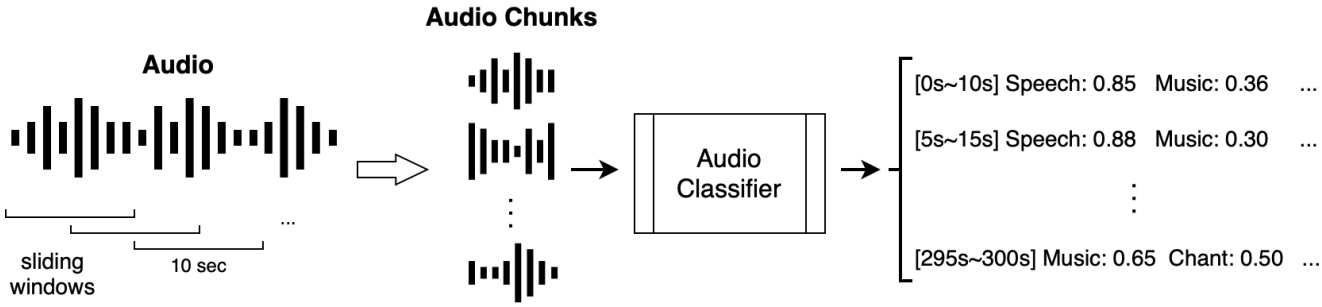

Figure 2. Pipeline overview for recognizing audio events in video clips.

in our main results. To be specific, StreamSync encounters a 45.2% degradation in performance when comparing between videos with talking heads and those in the Others category. The pretrained SparseSync has a more significant 54.4% degradation.

However, while these results indicate RealSync may induce less Talking Head bias, a 45.2% advantage for talking head situations still represents a significant performance bias. Future work should investigate methods for encouraging diverse audio event utilization to reduce this bias further. In these efforts, the detailed annotations provided by RealSync may be of significant utility and should be adopted by other datasets when possible.

References

[1] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022. 1

[2] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 251–263. Springer, 2017. 1

[3] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201, 2017. 1

No.	Scenario	Success/Failure	Evidence/Reason for Lack of Synchronization
1	Talking Heads	Success	Speech and lip movement
2	Talking Heads	Success	Speech and lip movement
3	Talking Heads	Failure	Presence of multiple individuals
4	Talking Heads	Failure	Presence of multiple individuals
5	Talking Heads	Failure	Periodic head movement
6	Voiceovers	Success	False negative from SyncNet (side profile)
7	Voiceovers	Success	False negative from SyncNet (obscured face)
8	Voiceovers	Failure	Confusion caused by voiceovers
9	Voiceovers	Failure	Confusion caused by voiceovers
10	Others	Success	Scene transition
11	Others	Success	Striking a golf ball
12	Others	Success	Hitting a tennis ball with a racket
13	Others	Failure	Smooth video events (musical performance)
14	Others	Failure	Smooth video events (applause and chant)

Table 1. Examples of success and failure cases, with explanations and evidence provided.

	Acc	Acc^{tol1}	$ROCAUC$	mAP
Talking-heads	0.258	0.465	0.772	0.226
Voiceovers	0.191	0.357	0.708	0.170
Others	0.119	0.252	0.622	0.103
Overall	0.201	0.378	0.717	0.174

Table 2. Evaluating the impact of talking heads presence in the scene for Baseline SparseSync, which was trained on LRS3 and VGGSound. Testing was done with 18 streaming iterations and 1 second hop size. These results show a 54.4% performance degradation between the talking heads and others categories. StreamSync, which is finetuned on RealSync, shows only a 45.2% degradation in comparison.