Supplementary Material: AVA-ActiveSpeaker: An Audio-Visual Dataset for Active Speaker Detection

Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, Caroline Pantofaru

Google Research
{josephroth, sourc, klejcho, radahika, agallagher, lkaver, sharadh, astopczynski, cordelias, zxi, cpantofaru}@google.com

Section 1 contains additional information on the dense, spatio-temporal labels in the AVA-ActiveSpeaker dataset that we will release and the labeling process, and Section 2 adds additional information related to model performance on this dataset. A more detailed treatment of the subject matter is also available in this arxiv paper.

1. Supplementary Dataset Information

Rating Interface: A larger image of the rating interface (Figure 2 in the main paper) is in Figure 1, with UI components highlighted. Raters can navigate using the timeline, modify labels, view tracks in the context of the video, and use keyboard shortcuts to view at varying speeds.

Videos with labels visualized: Two clips with labels visualized are linked below. The bounding box color indicates the label: red: “Not Speaking”, green: “Speaking and Audible”, and yellow: “Speaking but Not Audible”.

1. speaker-labeled-1: A “conversation” between 3 people. Near the end of this clip, the spoken content is replaced with music as an illustrative example for the “Speaking but not Audible” label.

2. speaker-labeled-2: Audible speech between the two participants. Note that one of them speaks (“Mark me”) before entering the visual scene, and that portion of speech is not part of the labeled data, since the speaker was not visible.

Release Data Format: The active speaker data can be downloaded from the AVA website. Figure 2 shows an example entry.

Rating Guidance: Detailed instructions for a variety of cases are summarized in Table 1. Speaking faces are audible if raters could ascertain that the subject’s speech was audible in the audio; clearly dubbed speech, speakers not heard due to overlapping sounds, or music (or other sounds) overlaid for cinematic effect is labeled “Speaking but not Audible”. The raters had access to both the audio and video while labeling the data.

AVA Action Label Differences The choice of the AVA corpus allows us to correlate ActiveSpeaker labels with previously-released action labels [24] (where annotators only had access to the visual modality) and speech activity labels [8]. Table 2 in the main paper notes that ~17% of “talk-to” labels and ~13% of “sing-to” in the AVA actions dataset were errors that do not correspond to someone speaking (audible or inaudible) in the ActiveSpeaker dataset. Below, we describe three conditions in the AVA actions dataset discovered by cross-checking the action labels with our speaker labels.

- Incorrect “talk-to” labels: Figure 3 shows instances where the face was not speaking but was labeled “talk-to”. These usually occur just outside the boundaries of a speaking segment. The AVA action annotators were misled by both the lack of audio in the labeling interface, and proximity to speaking frames.

- Inaudible “talk-to” labels: The true semantics cannot be ascertained from visual-only labeling: the face looks like its speaking, but it is not heard in the audio which often contains overlaid music for cinematic ef-
Figure 1. The annotation interface for AVA-ActiveSpeaker, with the interface components marked with yellow boxes. As rating progresses, the labels are overlaid on the audio waveform timeline - the snapshot here shows a fully labeled timeline. Raters can skip to any point in the video and the box around the subject’s face corresponds to the color of the label at that instant. Raters can modify labels as necessary.

Figure 2. Release CSV Data Format. The first line from “release_example_-IELREHX_js_1740_1800_5.csv” broken down by the eight comma separated values. \((x_1, y_1)\) and \((x_2, y_2)\) are the normalized locations of the top left and bottom right of the bounding box.

AVA Speech Label Differences: Section 4.2 of the paper discusses the AVA-ActiveSpeaker labels in the context of the previously released speech activity labels [8]. One surprising observation was that a significant amount of time labeled as containing speech in AVA-Speech did not have an active speaker at that instant in AVA-ActiveSpeaker (Figure 4 in main paper). Sampling such cases shows that the shot often does not directly focus on the active speaker in movies for cinematic effect; viewers have enough context and voice recognition to know the speaker, anyway. Background music with vocals are labeled as containing speech, but naturally cannot be associated with a visible speaker.

Incorrect “sing-to” labels: Figure 5 shows cases where the specified entity was not actually singing; once again, the absence of audio makes it hard for the annotators to accurately determine when someone is singing or just lip-syncing along with the music.

Overlapping Speakers: The detailed per-person labels allow us to identify segments with overlapping speakers, potentially interesting for audiovisual speech separation efforts. Figure 6 shows snapshots from such segments identified with the dense labels in AVA-ActiveSpeaker.

Missed Face Tracks: Section 3 of the paper describes the automated process for obtaining the face tracks which form the basis of the labels in AVA-ActiveSpeaker. While the detection and tracking pipeline is state-of-the-art, the videos in this dataset contain a number of challenging cases such as crowded scenes, small faces, challenging lighting conditions, partially occluded faces, etc., in conjunction with the varied resolutions of the video itself, many of them lower than the production quality of movies and TV shows of today. This leads to some missed face track detection which are not labeled. Figure 7 contains a sample of instances where faces were missed.

2. Supplementary Results

As in the main paper, V: visual-only model, AV: audiovisual model, GRU: gated recurrent unit models, \(f_M\): number of frames in the stack input to the visual network, TPR: True positive rate, FPR: False positive rate.

Model Comparison: Figure 8 contains the full ROC.
Figure 3. Frames with incorrect AVA action label “talk-to” when the subject (in the pink bounding box) wasn’t speaking.

Figure 4. Frames with “talk-to” AVA action label where the subject appears to be speaking but their speech is not in the audio track, instead overlaid with music or sound effects for cinematic effect.

Figure 5. Frames of “sing-to” AVA action labels where the subject was not actually singing. All examples have a musical context. (Left & Center) The subjects were dancing to music. (R) Subject is a conductor directing the musicians, with emphatic mouth (and body) motions.

Figure 6. Visualization of overlapping speaker instances. A green bounding box represents speaking and audible, yellow represents speaking and inaudible, red represents not speaking.

Figure 7. (Left) Undetected backlit faces; (Center) A few missed faces in the back of a crowd; (Right) Small faces are often missed.
curves for Table 5 from the paper. In static models, performance keeps improving till \( M = 10 \), and for each \( M \), the corresponding AV curve is considerably better than V; AV-GRU is \( \sim 10\% \) better TPR than V-GRU and \( \sim 5\% \) better TPR than AV-static at 10% FPR. The same pattern holds with recurrent models, although performance improvements saturate at 2 frames, indicating that only a short amount of history is needed.

**Effect of background noise:** Figure 9 shows the full ROC curves for Table 6 from the paper. Unlike audio-based speech detectors’ performance reported in [8], both V and AV models show resilience to background sound. While V models are not affected at all, AV models’ performance slightly dips in overlapping music and noise, although still outperforming V models.

**Effect of face size:** Figure 10 shows ROC curves for Table 7 of the paper, partitioned by face size: small (\(< 64 \) pixels wide), medium (\( > 64, < 128 \) px) and large (\( > 128 \) px, larger than model input). AV models clearly outperform V models: for GRU, absolute improvement in TPR at 10% FPR is \( \sim 10\% \) for small faces, \( \sim 15\% \) for medium, \( \sim 13\% \) for large. The biggest difference in “medium” suggests that this might be the sweet spot for the combined advantage of recurrence and AV: for smaller faces, the visual information is harder to leverage for all models, while for larger faces, visual information is enough for V to close the gap.
For AV models, FPR at the balanced accuracy point is nearly constant, while for V, they are more variable. For applications that mine data corresponding to speaking faces (for tasks like synchronization [9], visual speech recognition [41], enhancement [1]), these models can be deployed without needing additional calibration and hand-tuning.

**Examples of model predictions:** Figure 11 shows frames where V model made errors while AV were correct. AV models appear more robust to pan angles and profiles (left two panels), can use audio context to know speech isn’t occurring (second from right), and are more robust to partial occlusions and motion around the face (right panel).

Figure 12 shows frames where AV static models were wrong but AV GRU got them right. Improvements from static to GRU within AV models appear to be driven by an enhanced ability to understand synchronization between the audio and visuals, even though the models were not explicitly trained for it. This makes the better models robust to noise in the audio domain (background music) as well as visual domain (partial occlusions).

Figure 13 shows frames where AV-GRU models made the wrong prediction. Based on our sampling, there appear to be two clear modes of failure. One occurs when the faces are small and there is motion in multiple faces (left two panels), where the model will pick both as speaking possibly due to not having a clear enough visual for who to associate the speech with. The other is when multiple sources are vocal but only one is speaking and others are, e.g., laughing (right two panels). One way to alleviate these issues would be to add in augmentation at training time geared toward enabling the model to learn explicit synchronization.