A Speaker-Resolved Multimodal Dataset of CEO Public Appearances, 2018–2023

Extended abstract submitted to the NBER Innovation Information Initiative (I3) Technical Working Group, December 4–5, 2026

Mine Ertugrul, Karthik Krishnan, Qianqian Yu

1. What we propose to present

We propose to present a dataset created using an extraction pipeline that convert publicly posted video of CEO appearances into a structured, speaker-resolved panel of visual, vocal, and textual measures linked to standard firm identifiers.

The dataset, in our new paper, "**Seeing Is Believing? The Information in CEO Visual Cues,"** covers 14,191 fully processed videos featuring chief executives of ExecuComp firms between January 2018 and December 2023, drawn from an identified universe of 17,326 videos across 6,048 distinct publishing channels. To our knowledge it is the first large-scale multimodal corpus of executive communication assembled outside the earnings-call setting; the closest comparable study in the literature analyzes 363 earnings videos.

The distinguishing feature of the resource is that every measure is constructed at the speaker level rather than the video level. Video, audio, and transcript are synchronized through diarization so that the facial expressions, vocal characteristics, and language attributed to the CEO are separated from those of interviewers and other participants. This makes it possible to hold the interviewer's contribution fixed, to benchmark an executive against the person questioning them within the same recording, and to control textual and vocal content when isolating visual information — none of which video-level aggregates support.

Our extraction pipeline uses an ensemble of models, mostly open source except for tools and models required for classification where we use Anthropic's models for higher quality of output. We use relatively cheap hardware for processing, which creates a longer processing time than we anticipated.

2. Sampling frame and video collection

The sampling frame begins with the universe of Compustat firms covered by ExecuComp, which reports executive-level information for S&P 1500 firms and other firms filing proxy statements. We identify the CEO of each firm using the ExecuComp annual CEO indicator and construct a panel at the CEO–firm–year level for fiscal years 2018 through 2023, yielding 10,626 CEO–firm–years covering 2,998 CEO–firm pairs and 2,031 firms. The start date reflects a sharp decline in video availability before 2018, and the end date leaves room to measure longer-horizon outcomes.

For each CEO–firm–year we search YouTube once per calendar quarter of the executive's tenure, using the standardized firm name together with the CEO's first and last names and restricting results to videos published in that quarter. A key hurdle for us is YouTube has stringent rate limits which constrain search even very small video samples; another is video downloads for data extraction are also limited based on rate and IP limits imposed by YouTube. We partially get over these hurdles by getting access to higher API limits from YouTube directly for our research.

Quarter-by-quarter searching is a deliberate design choice: unrestricted searches return results ordered by platform relevance, which concentrates coverage among highly visible executives and firms. We retain only videos in which both names appear in the title or description, and we restrict the window to the period during which the executive served as CEO of the focal firm.

Table 1 summarizes attrition. Of the 14,440 videos submitted for processing to date, 242 could not be downloaded, almost all because they had been made private or removed between search and retrieval, and two failed during processing. Processing of the remaining 2,886 identified videos is in progress.

Table 1. Sample construction

Step	Observations
ExecuComp CEO–firm–years, FY2018–2023 (2,031 firms)	10,626
Unique videos returned by quarterly search and passing name filter	17,326
Distinct publishing channels represented	6,048
Videos submitted to the extraction pipeline to date	14,440
Videos completing every extraction stage without error	14,191
Video events merged to CRSP and ExecuComp (1,001 CEO–firm pairs)	10,881
Video events with complete controls used in reported tests	5,603

3. Extraction pipeline and speaker resolution

The pipeline proceeds in three stages. First, audio is transcribed with timestamps using Whisper (large-v2), and speaker turns are identified independently using the Pyannote diarization model; the two are aligned on timestamps so that each transcript segment carries a speaker index. Second, because diarization distinguishes speakers without identifying them, a large language model (Anthropic Claude Sonnet) reads the diarized transcript together with the CEO and firm identifiers and labels each speaker as the CEO, the interviewer, or another participant. Third, the resulting CEO-specific intervals are used to cut the corresponding transcript text, audio, and video frames, from which the three families of measures are built.

Facial measures are produced with py-feat, an open-source toolbox that uses MTCNN for face detection and PFLD for landmark detection and classifies seven expressions: happiness, anger, fear, sadness, disgust, surprise, and neutral. Because frame-by-frame processing of the corpus is computationally infeasible, we sample three frames per speaker per video from candidate points at the beginning, midpoint, and end of each speaking segment, discarding frames in which no face is detected. Classifications are binary at the frame level and are aggregated to the speaker–video level as the share of analyzed frames in which each expression is detected, so the measures capture how often an expression appears rather than its intensity. We release all seven expression shares rather than only composite indices, so that users can build their own aggregations.

Vocal measures are computed from the speaker-specific audio; the primary measure is shimmer, the short-term variation in vocal amplitude across successive vocal-fold vibrations. Textual measures are computed on the concatenated segments assigned to a speaker and include VADER compound sentiment and its component shares; TextBlob polarity and subjectivity; hedge-word and confidence-word ratios and their difference; and delivery-style measures built from the spaCy pipeline, including the share of sentences that are questions, the first-person pronoun ratio, the incidence of passive constructions, the ratio of authority words, and average words per sentence. Because the interviewer's speech is processed identically, all of these can be differenced within a video.

The release also includes interview-level metadata that should be reusable independently of the multimodal measures: non-exclusive indicators for 23 topics under six headings (earnings, strategic announcements, crisis management, financing activity, macroeconomic and industry commentary, and reputation and personal branding); interview format (one-on-one, panel, award event, other) and setting (studio, remote, unknown); counts of total speakers and of speakers affiliated with the focal firm; and a channel-level classification of all 6,048 publishers as corporate promotional, business and finance news, association or institutional, other, or unknown, with an indicator for major business news outlets. Each video event carries the identifiers needed to merge with ExecuComp, Compustat, and CRSP.

4. Validation: do the extracted measures carry signal?

We report results from our own work only to establish that the measures behave in economically sensible ways and are not dominated by classifier noise. Visual measures are standardized, so coefficients are per standard deviation. In event-time regressions of cumulative abnormal returns on the visual measures, controlling for vocal and textual characteristics, CEO and firm characteristics, interview topic, and industry-by-quarter, interview-type, and interview-setting fixed effects, a one-standard-deviation increase in the positive visual measure is associated with roughly 20 basis points of abnormal return over the [0,1] and [0,3] windows, significant at the five percent level under both Fama-French and market-model benchmarks; the composite negative measure is indistinguishable from zero. Over longer horizons the pattern reverses: the positive measure has no predictive content, while a one-standard-deviation increase in the negative measure is associated with 0.6 percentage points lower buy-and-hold abnormal return at three months and 1.4 percentage points at six months. The negative measure also predicts lower return on assets two quarters ahead and on average over the following four quarters, and results are unchanged when the common-factor construction is replaced with principal components.

For present purposes the point is narrower than the economics: measures built from three sampled frames per speaker, after conditioning on what the executive said and how they sounded, retain predictive content for outcomes realized months later.

5. Value to the research community

- Innovation and opacity. The predictive content of the visual measures is concentrated in high R&D-intensity firms, where conventional disclosures are least informative about future cash flows. Researchers studying how innovation-intensive firms communicate with capital markets have few measures of what is conveyed outside formal disclosure; this dataset supplies several thousand such events per year.
- Executive communication and disclosure. The corpus covers an unstructured, largely unstudied disclosure channel. Unlike earnings calls, interviews are not scheduled, not scripted, and not confined to financial results, and the topic and setting codes allow researchers to condition on what is being discussed and where.
- Media and information intermediaries. Channel-level classification of 6,048 publishers, combined with publication dates and topics, supports work on which outlets cover which firms and how executive access is allocated across media.
- Methodological reuse and benchmarking. The pipeline is agnostic to who is speaking: applied to a different sampling frame it yields comparable measures for founders pitching, scientists presenting, or officials testifying. Because we retain intermediate output — diarization boundaries, speaker role assignments, and frame-level classifications — the corpus also serves as a testbed for comparing alternative diarization, speaker-attribution, and expression-classification models on in-the-wild rather than curated recordings.

We will document the limitations users should carry into their designs: three frames per speaker is a coarse sample; binary expression classifiers measure facial configurations conventionally coded as positive or negative rather than internal emotional states; platform search yields coverage that is not random across firms; and videos disappear, which makes the corpus non-reproducible from the search alone.

6. Dissemination plan

We plan a staged release tied to the paper's progress through the publication process.

Stage 1, at working-paper circulation. We will release summary statistics for the full set of extracted visual, vocal, textual, and interview-level measures; platform video identifiers for every video in the corpus; and diarized segment boundaries for each video, with speakers indexed but not assigned to roles. This makes the segmentation layer — the output that is most expensive to produce and most reusable in

other settings — available immediately, while speaker attribution and the derived features will be distributed after the publication process

Stage 2, on publication/acceptance. We will release the full derived dataset: video-event identifiers, publication dates, CEO and firm identifiers, speaker-resolved visual, vocal, and textual measures, and the interview and channel classifications, distributed as flat files with a variable dictionary and a linking table to ExecuComp and Compustat identifiers. Because the recordings are third-party copyrighted works, we will not redistribute video or audio at any stage; users retrieve what remains available using the identifiers and segment boundaries.

Stage 3, on publication/acceptance. Alongside the data we are considering releasing the extraction pipeline as a documented open-source repository under a permissive license, covering video retrieval, transcription, diarization, transcript–diarization alignment, and feature extraction across the visual, acoustic, and textual channels. Model versions and parameters will be pinned so that output is reproducible conditional on the retained inputs. The release will be hosted in a versioned archive with a citable persistent identifier and an issue tracker for corrections.

We are currently considering setting up a public access API to allow researchers to directly diarize and extract data videos that they submit. Depending on demand, we will first need to ascertain availability of computing resources to allow for such a service.

7. Focus of the presentation

A deep dive would concentrate on the speaker-resolution stage, which most affects downstream validity and is the part most transferable to other settings, and on the design choices that shape coverage. We would also use the session for feedback on the questions we are least settled on: whether to release intensity-weighted as well as binary expression measures, how to document language-model-based classification, and how to structure the release so that others can extend the corpus to speakers beyond public-company CEOs.