Human Interactions and Financial Investment: A Video-Based Approach*

Allen Hu Song Ma

Economic decisions are often made after human interactions. This paper proposes a flexible and extendable empirical framework to study micro-level human interactions and their connections to financial investment decisions. We project human interactions to three-V dimensions—visual, vocal, and verbal, and construct easily interpretable metrics along these dimensions using machine learning algorithms with videos as data input. The framework and methodology are applied to pitch videos in which entrepreneurs pitching investors for funding. We find that startup founder(s)'s behaviors differ significantly in their pitches. Venture investors are more likely to invest in startup teams that show more positivity (i.e., happy, passionate), but startup teams showing such features underperform conditional on investment. These features in human interactions carry more weight when evaluating female entrepreneurs individually, but women are often neglected when presenting together with their male teammates. We do not find strong evidence supporting that human interactions allow venture investors to extract additional valuable information about entrepreneurs on net. In contrast, these interactions amplifies social biases.

JEL Classification: D91, G41, G24, L26

Keywords: Social Economics and Finance, Video Data, Machine Learning, Entrepreneurship

*Hu and Ma are both with Yale School of Management. Emails: allen.hu@yale.edu; song.ma@yale.edu. We thank Nick Barberis and Yueran Ma for many useful conversations. We also thank Xuechunzi Bai, James Choi, Cydney Dupree, Mike Ewens, Susan Fiske, Will Goetzmann, Paul Goldsmith-Pinkham, Arpit Gupta, Kyle Jensen, Bryan Kelly, Michael Kraus, Kelly Shue, Olav Sorenson, Alexander Todorov, Alex Zentefis and seminar participants at Colorado (Leeds), Georgia State (Robinson), Indiana Kelley Junior Finance Conference, Washington St. Louis Corporate Finance Conference, and Yale for helpful comments. Ran You provided excellent research assistance. All errors are our own.

1. Introduction

Many economics decisions are made after pitches, in which one motivated agent (signal *sender*) persuade the other party (signal *receiver*) to make certain decisions in an interaction. For example, corporate executives pitch the board for certain corporate decisions, sell-side analyst market product to the buy-side, startup entrepreneurs make pitches to potential investors to seek funding, or broadly, politicians speak at campaign events to engage potential voters. It is widely believed that in those personal interactions, not only does the content matters, but also the features of delivery itself. These features, both verbal and nonverbal, could be informative about the quality of the object (a product, a startup) under consideration, but could also introduce forces rooted in social psychology.

Accompanied with the agreement on the importance of personal interactions is the equally strong needs for answers to some fundamental empirical questions. What is the empirical structure of human interactions—that is, how can we empirically represent, measure, and study the features from such interactions? What features in human interactions matter for economic decisions, and why? Are human interactions beneficial to economic decision making, or just a sources of introducing and amplifying social biases?

This paper takes a step toward answering these questions using field data and a methodological framework new to economics. First, our primary contribution is methodological. We propose to use video as data input and develop a machine learning-based framework to process videos and extract pitch features (e.g., facial emotions, audio valence, etc.). This empirical framework is highly flexible, extendable, and easy to scale up, and can help with many relevant questions. Second, we apply the methodology to a setting of startup companies pitching for venture funding. We show that the way people convey information, often nonverbal, closely relates to the probability of receiving funding. Higher positivity and warmth in pitches are associated with a higher probability of receiving venture funding. However, these features do not imply higher quality, as they negatively correlate with future performance. In contrast, their impacts vary in ways consistent with gender stereotyping and first and superficial impressions.

Let us get into the basic structure of the methodology and our application of startup

pitches. We focus our study on a setting of startup founder(s) pitching to early-stage investors, and use those pitch videos as the unstructured data input. In specific, we collect thousands of highly standardized pitch videos of startup founders at the time of applying to top accelerator programs. Those videos are roughly one to two minutes long in which a startup founder(s) pitch the entrepreneurial idea. We observe the accelerator investment decisions and future performance of these firms. This design captures the conceptual experiment in which investors observe entrepreneurial pitches and then make investment decisions.

The next hurdle concerns how to analyze unstructured video data and extract features of how startup founders pitch. We argue that for the purpose of economic studies, pitches can be projected onto three-V information spaces, namely the visual (e.g., facial expressions), vocal (e.g., tone of voice), and verbal (e.g., the content of communication and sentiments).[1] These information is worth studying as standalone variables, but could also be studied jointly. Imagine a startup founding team is making a pitch to VC investors about their startup idea. The three-V structure is analogous to how venture capitalists perceive the pitch. The information includes the idea of the startup and specific word choices when describing it (verbal/textual), how they say it as reflected in the features of the voice (voice/vocal), and the facial expressions and other potential body languages in the presentation (visual).

To implement this empirical structure, the next step is to clearly capture information along those dimensions using the pitch video data. We exploit state-of-the-art machine learning algorithms and develop a set of tools to extract information carried by the videos. Specifically, we represent audio continuously, represent video as a series of images at the frequency of 10 frames per second or more (when the video is with high resolution),[2] and extract the full textual script of the pitch using speech-to-text algorithms.

Then we construct interpretable measures from these information channels with established techniques of facial and vocal emotion analysis and dictionaries coded by researchers in economics and social psychology. For the purpose of replicability and transparency, we use easily-accessible computation services, such as Face++, Microsoft Azure, and Google Cloud,

[1] This paper will use the following pairs of terms interchangeably depending on context: verbal and textual, vocal and voice, visual and facial.

[2] This is a very small time interval for the purpose of our study and the goal of capturing facial expressions. Even the fast facial movements, like blinking, takes on average around 0.25s, and will be captured by our algorithm at the 1/10th second intervals.

to perform the computation. These measures include visual emotions (e.g., positive, negative) and beauty, audio emotions (e.g., positive, negative, valence, arousal), and textual sentiment and psychological features. These information is then aggregated to a startup/video-level data. The final analytical dataset is essentially a firm/video cross-section, for which we are able to observe the aggregated measures. We use a principal components analysis to construct a simple measure of "Pitch-Factor" that captures the positivity in the videos.

The measures that we focus on complement to the subject-rating-based measures of perception and personality (Kaplan et al., 2012; Graham et al., 2016; Blankespoor et al., 2017, for recent examples). The machine-learning based method presents several advantages. First, the measures give us the opportunity to separately examine the effects of visual, vocal, and verbal channels, which is difficult to achieve with rating-based measures. Second, these algorithm-constructed measures are more objective and free from employing laboratory subjects, who may form biased perceptions themselves. Third, the process is highly reproducible and is ready to be scaled up easily. All those advantages allow for future research to use video data of different forms and in creative ways.

Armed with the measures, our analysis proceeds in three steps. To set the stage, we show that there are important variations when characterizing entrepreneurial pitches. Even though the video pitches are only one and half minutes long on average, the extracted features vary along all three-V dimensions. For example, based on the Visual-Positive measure, entrepreneurial teams at the 25th percentile show their happy motion in only one third of the time than teams standing at the 75th percentile. Based on the Vocal-Arousal and Vocal-Valence measure, pitches at the 25th percentile shows only half the passion (energy and positivity) compared to ones at the 75th percentile. In addition, we also show that features represented across vocal and visual dimensions are only mildly correlated, suggesting the independence of those information channels.

Next, the baseline results show that the pitch features, along all three-V dimensions, have a strong correlation with the investment decision of venture investors. Startup teams that score higher in "warmth" and "positivity" dimensions are more likely to be invested, and this is consistent across measures from vocal, visual, as well as textual dimensions. Using the single factor Pitch-Factor, we find that a one-standard-deviation change is associated

with 35.2% increase in the probability of the startup being invested. This applies to detailed specific features too. For example, a one standard deviation change of positivity in the visual (vocal) dimension is associated with a 17.6% (10.6%) increase of the probability of receiving accelerator funding. Meanwhile, startup teams who lacks such features in the pitch are less likely to be invested, and the economic consequence of behaving negatively is more sizable. A one-standard-deviation increase in visual (vocal) negativity is associated with a 31.7% (52.8%) decrease of the probability of obtaining accelerator financing. Even though we find the "beauty premium" (Hamermesh and Biddle, 1994), in the context means that more beautiful (ML-generated) founders are more likely to be invested, the dimension works independently from features of the pitch.

Why do pitch features matter for investment decisions? We focus on investigating two potential channels. First, pitches may be good signals of the objective startup quality. For example, positivity features may capture higher startup quality that are otherwise hard to measure. This could be either the ability of entrepreneurs, or information available to only entrepreneurs (if better startups behave more positively). Second, pitches allow investors to develop (possibly biased) perceptions and impressions about the entrepreneur(s). Positive pitch features may lead investors to perceive a more positive impressions on the entrepreneurs, which fits startup teams into the successful entrepreneur stereotype often characterized as optimistic, energetic, and ambitious.

We investigate the "quality" hypothesis in two ways. First, we add quality controls to the investment regression, including hand coded information on the education, employment background, and startup experience of the founders. The results are statistically robust. But more importantly for us, the coefficients remain economically stable in magnitude. Based on Altonji, Elder, and Taber (2005) and Oster (2019), this stability of the coefficients means that the pitch features are not crowded out by quality information, suggesting that the quality component in these features is minimal.

Another test to investigate whether pitch features incorporate startup quality is by studying the long-term performance of the invested startup. Under this channel, the invested companies with higher level of positivity features should perform better than those without. We track the performance of startup using employment, and find that none of the positive

pitch features positively links to the long-term performance of the firm. In fact, many of them are negatively associated with startup performance. This suggest that, instead of capturing quality, those features lead investors to lower their bar when picking investments.

The analysis then switches to investigate the social psychology. Our first test builds on the idea that if social cognition is why pitch features matter, the finding should vary with subsamples where such forces behave differently. One of such subsample cuts is by the gender of the entrepreneurs. Previous research shows that women are often perceived differently in social interactions due to gender stereotyping or gender inequality (Brooks et al., 2014). Women are more often judged based on appearance and not substances (Fredrickson and Roberts, 1997). Women and men are also expected to follow different gender stereotypes—there are generally higher expectations of men in general ability and task performance domains, while women are expected to be high in terms of warmth, empathy, and altruism (Kite, Deaux, and Haines, 2008; Ellemers, 2018).

We find that the relation between pitch features and venture investment decisions differ across women and men entrepreneurs. When investigating man-only and woman-only teams, woman entrepreneurs are evaluated along similar dimensions and in the same directions as the man entrepreneurs, but with different intensity. Investment decisions are more sensitive to the level of positive attitudes among female entrepreneurs than among male entrepreneurs. Moreover, we find that women are often overlooked in such human interactions—an implication of the unpleasant social psychological bias. In mixed-gender teams, i.e., teams with both women and men, when putting the features of both the female and male speakers in the analytical framework, features of the male speaker remains similar as before, while that of the female speaker turns in significant. In fact, in these mixed-gender teams, investment probability increases with women behaving less active and more obedient.

In our last set of test, again in support of the role of social psychology, we show that the pitch features affect venture investment decisions through first impressions and superficial impressions. The first five seconds of the pitch video, which typically only includes a short greeting from the team, provide sufficient explanatory power for the investment decision. Beyond this first impression, we also directly show that investors do not actively exploit deeper business information from the pitch. Specifically, we show that investment decisions

do not form perceptions after taking into account the content of the pitch.

The remainder of the paper proceeds as follows. Section 2 provides motivation for studying the impact of personal interactions in the video form. Section 3 introduces data collection and processing. Section 4 describes the method of constructing interpretable metrics using machine learning technologies. Section 5 presents empirical analysis on human interactions and financial investment decisions. Section 6 discusses connections with literature, potential extensions, and concludes.

2. Theory and Background

One of the fundamental questions in financial economics is to explore factors that affect economic agents' decision making. Some well-studied factors include hard information (e.g., the expected return and volatility of an asset), agents' psychological and cognitive traits (e.g., overconfidence, limited attention), and cultural and social forces (e.g., cultural norms, peer effects), among others. Why do personal interactions, reflected in visual, vocal, and verbal features, matter for economics decisions? Potential explanations can be categorized into two mutually non-exclusive streams.

First, the investment decisions' sensitivities to human interactions may simply reflect the impact of social psychological forces. Leading social psychology theories agree that economic decisions are often based on the perception of the others (Fiske, Cuddy, and Glick, 2007; Cuddy, Glick, and Beninger, 2011). A natural consequence of these social psychology forces is the room for potential biases—inaccurate first impressions, stereotyping, halo effect, among others. These social cognitive biases are salience and impactful.

Relatedly, this stream also connects an active literature concerning the effect of personal traits and personality on economic outcomes. Kaplan, Klebanov, and Sorensen (2012) provides direct empirical evidence and show how personal traits matters for the selection of CEOs and what abilities matter for firm performance conditional on being selected. Malmendier and Tate (2005), Malmendier and Tate (2008), and Malmendier and Tate (2015) empirically show that CEO overconfidence is related to a wide set of corporate decisions and outcomes. Meanwhile, personal interactions are important for people to form opinions on people's

personalities (Knapp, Hall, and Horgan, 2013).

In both literature of this social psychology streams, face-to-face personal interactions are an important channel through which perceptions and judgement about personality and personal traits are formed (McArthur and Baron, 1983; Murphy, Hall, and Colvin, 2003; Adams, Adams Jr, Ambady, Shimojo, and Nakayama, 2011). Facial expressions, vocal features, and the choice of words in communication can impact economic decision making through influencing the process of perception formation and judgment on personality.

An alternative view, which is more neoclassical, argues that human interactions may bring additional information useful for the decision making. In a world with incomplete information, characteristics reflected in interactions and communications may serve as a sign of startup quality that are privately known to the startup founder. For example, people are happier due to higher quality ideas. Those characteristics could also matter due to the risk of running into a state of renegotiation and collaboration, say investors value the possibility to work with more positive and open minded people in difficult times.

Certain characteristics embedded in personal interactions could also be productive. For example, more socialablea and passionate people might be able to obtain more resources. Hamermesh and Biddle (1994) and Graham, Harvey, and Puri (2016) show that personal appearance (more or less beautiful) is associated with labor market outcomes, even at the CEO level, partially attributable to the productivity enhancing channel. Mobius and Rosenblat (2006) show that beauty is associate with different confidence levels and oral skills in negotiation.

Venture investing is an ideal setting to explore the relation between human interactions and financial decisions, and to apply our methodology. Venture investment faces significant information asymmetry (Kaplan and Strömberg, 2004; Kaplan and Stromberg, 2001), and one of the fundamental questions that are still not well understood is "How do VCs make their decisions?" Field evidence through surveys (Gompers, Gornall, Kaplan, and Strebulaev, 2019) and experiment-based evidence (Bernstein, Korteweg, and Laws, 2017) suggests that investors heavily value teams when making the investment decisions. Pitches are crucial in the investment decision making (Brooks, Huang, Kearney, and Murray, 2014; Huang and Pearce, 2015; Kanze, Huang, Conley, and Higgins, 2018).

3. Data and Setting

Our data set consists of two parts—entrepreneurs' pitch videos for accelerator applications and startups' firm- and team-level information. Two datasets are merged by firm names accurately by hand. The sample spans the period from 2010 to 2019.

A. Video Data

We focus on entrepreneur team self-recorded pitching videos in their applications for five large incubators in the U.S. These incubators are Y Combinator, MassChallenge, 500 Startups, Techstars, and AngelPad. As of July 2019, these incubators, letting alone MassChallenge, have accelerated 3,976 startups, which in total have raised over 48 billion dollars of total funding, taking up 76% of all accelerator backed companies and 97% of funding by all seed accelerators around the world.[3] Y Combinator alone, since 2005, has funded over 2,000 startups with a combined valuation of over 100 billion dollars. Some well-known examples are Dropbox (2007), Airbnb (2009), Docker (2010), and Doordash (2013).[4] MassChallenge, a zero-equity startup accelerator, has funded 2,344 startups and raised over 5 billion dollars of funding since its establishment in 2009.[5]

These incubators require (or highly recommend) each team to record and submit a standardized self-introduction video as part of the application process. Figure 1 shows such examples from those incubators' application system. These videos are typically one minute long (up to three minutes in our sample), and they present the founder(s) introducing the team and describing the business idea. These videos are highly standardized, and have several desirable features for the data processing described in the next section. For example, background music are forbidden; the setting is stable; faces are highly visible; and there is no scene-switching in the video.

[Insert Figure 1 Here.]

These videos, rather than being submitted to the incubators themselves, are submitted

[3]Statistics are from Seed-DB https://www.seed-db.com/accelerators.
[4]Information comes from Y Combinator's website https://www.ycombinator.com.
[5]Information comes from MassChallenge's website https://masschallenge.org.

through uploading to a public multimedia platform such as YouTube and providing the webpage links to these video in application forms.[6] We developed an automatic searching script for two public video-sharing websites, YouTube and Vimeo. Integrated with query APIs, our web crawler returns a list of video indices according to a lot of pre-defined keywords, which include but not limit to the names of these accelerators, and a list of keywords including "startup accelerator application video","accelerator application videos", among others. We describe the process of obtaining and filtering videos in detail in Appendix.A.

[Insert Table 1 Here.]

In total we obtain 1,139 videos. Table 2 describes the sample, in which the number of videos are reported by incubator (Panel A) and by year (Panel B). Y Combinator contributes the largest number of application videos, followed by MassChallenge and Techstars. Among all the companies that applied, 97 (8.52%) were chosen by the accelerator program, and 248 (21.77%) were invested by any venture investor (incubator or angels/VCs). The videos are more available for recent years due to the increase in video requirements in the application.

After collecting the videos, we parse each video webpage to collect other relevant information. This includes the video's duration, upload date, title, annotation, subtitle, and uploader ID. This set of information also allows us to identify the startup almost perfectly. Specifically, by scrutinizing video titles and annotations, we double check names of the startups and names of the accelerators they are applying for. If the startup name cannot be identified from these items, we search the uploader name on LinkedIn and back out the company information. It is common that many people have the same name on LinkedIn, to verify that the person on Linkedin is the founder, we also double check the name, background, experience, and even photos.

[Insert Table 2 Here.]

In Table 2 Panel A we report basic information at the level of video pitches. The median length of a pitch videos is 68 seconds. These videos are not viewed often and do not attract

[6]For detailed information, please check Y Combinator's video guidance. https://www.ycombinator.com/video

almost any likes or dislikes. This is consistent with that those videos are generic pitch videos for application purposes rather than for any marketing campaign or product promotion. Regarding the team composition, we find that 46% of the startups have only one member pitching, and 54% have multiple members. The average member per video is 1.74. 49% of the teams have only male founders, 27% have only female founders, and 24% have mixed genders.

B. Startup Information, Venture Financing, and Team Background

We collect firm-level information from two widely-used entrepreneurship databases, Crunchbase and PitchBook. We start by searching firms in these two databases according to the names identified in application videos using web crawler. Our script sends requests of startup names to the search engine of Crunchbase and PitchBook. Then URLs of the startups are returned if the firms are included in these two datasets. For startups with duplicate names, we identify firms by names of founders, business descriptions, and time of founding.

We then iterate all these company profile links and download the text content from the web. The HTML files are parsed to extract firm-level variables including the year of founding, location, operating status, total funding round and amount, number of investors and employees, and other information from the main profile page. We also hand collect more detailed information on firms' funding, deal history, partners, team and board members, similar companies and competitors using sub-page links for each firm. We also check the startups' "names used before", "also known as", and latest names. If a startup indeed changes its name or has multiple names, we try all its alternative names in the matching process.

Beyond firm-level information, we also collect information on the founding teams. We compile a list of founders for each startup company using PitchBook, Crunchbase, and video content. We also design an algorithm that can automatically search company names on LinkedIn and find the person with the most senior position of each company. We used this algorithm, assisted with manual check, to retrieve individual profiles from LinkedIn. For each startup team member, we extracted 5 most recent education experiences and 10 most recent work experiences from each profile. We then used the information to construct variables that indicate her/his education at elite university, degree level, job seniority, entrepreneurship

experience, etc. Furthermore, we also code the race of founding team members for each company. For each individual in application videos, we manually code her/him into one of the five racial groups–White, Black, Hispanic/Latino, Asian, Others. We also standardize firms' industry classifications using Global Industry Classification Standard (GICS). By referring to the industry information in Crunchbase and PitchBook along with the video scripts, we categorize all companies into one of 24 GICS industry groups, which are then grouped into 11 GICS sectors.

Finally, we merge our video information dataset with the firm dataset. Among these 1,139 applications, 462 are included in Crunchbase, and 217 can be found in PitchBook. The summary statistics of those startups are reported in Panel B of Table 2. On average each startup, conditional on being invested by at least one investor, raises $11,804,000 in total, and the median is $382,000. The median financing round is one and the number of employees at the time of data collection is five. Regarding the startup founder teams' backgrounds, we find that 30% of the founders have startup experience at the time of application. 19% and 3% of them hold master and PhD degree, respectively.

4. Method: Processing Video Data with ML Algorithms

This study aims to develop the methodology of quantifying micro-level human interactions using video data. Video data truthfully record human behaviors—how people look like (visual, i.e., facial expressions and emotions), what people say (verbal, i.e., text), and how they say it (vocal, i.e., tone of voice, vocal emotion). However, coming with this richness is the fact that video data are highly unstructured, high-dimensional, highly noisy, all of which impose technical difficulty in information extraction and a substantial computation burden.

Our method proceeds in three steps. First, we decomposes the information embedded in videos to three-V dimensions—visual, vocal, and verbal. Second, for each dimension, we adopt machine learning algorithms to create interpretable visual, vocal, and verbal features from the raw data. Third, we aggregate these measures within and across these dimensions to characterize features presented in each pitch video. This framework is flexible and extendable for more complicated video data than pitches. Below we first overview the key properties of

video data relevant to economic research, and then describe the three-step methodology in details.

A. Video as Data: Key Properties

Let us begin by discussing some basic properties of videos and illustrate how these properties relates to our video processing and measure construction procedure.

First, video data are information intensive. In our settings, this means that videos can deliver diverse aspects of human behaviors and high-fidelity details of these behaviors. In other words, video data is inherently multi-modal and can comprehensively characterize human interactions. To better understand the richness of video data, consider the following comparison. The csv-format startup panel in this study is around 1MB large. It includes 150 firm-level characteristics collected from Crunchbase and Pitchbook for around one thousand startups. In contrast, a one-minute high-resolution pitch video in the mp4 format can be as large as 200 MB in our sample. To put this into perspective, one second of a high-definition video, in terms of size, is equivalent to over 2000 pages of text (Gandomi and Haider, 2015; Manyika et al., 2011).

Second, video data are unstructured and high-dimensional, making them more complicated to process relative to other data formats such as panel data, or even unstructured textual data. Consider a one-minute video with a resolution of 1280×720 (720p) and two 48 kHz audio channels. The resolution parameters mean that there are 1280×720=921,600 pixels in each frame of this video. If we employ a sampling rate of 10 frames per second, the video can be represented as a series of 600 static images. This means that we need 552 million pixels (dimensions) to represent the visual information in just one video. Further, to represent the acoustic information, we need 48,000×2 dimensions per second and thus around 5 million dimensions for one minute. In total, to represent such a video digitally, we need around 560 million dimensions.

Third, video data are low in signal-to-noise ratio (SNR) and in economic interpretability. Regarding the SNR, most videos have a large amount of background noises that are likely irrelevant to the economic question researchers study. In our setting, these noises include background environment noise, furniture in the video, among others. Moreover, the unit

of data in a video, say the pixels, the acoustic features, etc., are not directly interpretable as accounting variables or textual words. Thus, when processing video data, we need to impose structures and algorithms to guide the extraction of information that are useful and meaningful for economic research purposes.

B. Information Structure and Representation of Video Data

The first step in video data processing is to project videos on a three-V structure and to represent them in a data format. Three-V means visual, vocal, and verbal; and this structure is intuitive and widely accepted in social psychology and communication literature (Mehrabian, 1972; Strahan and Zytowski, 1976; Krauss et al., 1981; Knapp et al., 2013).

We first extract these three dimensions from the two streams of digital information in raw video records—the image stream and the sound stream. To be more specific, in our setting of studying human interactions, facial expressions and speeches are embedded in the streams of image and sound, respectively. For visual, we represent the video using images sampled at 10 frames per second and employ face detection machine learning algorithms to identify human faces in each video frame. For verbal, a speech-to-text machine learning algorithms is used to extract speech from the sound. For vocal, we extract the audio files from the video.

We now consider the features of these three dimensions and discuss their data representations one by one. For visual, the portraits of human faces exist in the format of digital image, which can be numerically represented as two-dimensional matrices. And since we have a stream of such digital images, the information in the facial dimension is coded as a time series of two-dimensional matrices. The vocal signal, essentially sound waves, can also be digitized as time series of amplitude given a specific sampling rate (the number of times per second that the amplitude of the signal is observed) and resolution (the number of discrete levels to approximate the continuous signal). If the audio is multi-channel, then it can be represented as a matrix with channels as columns/rows. Finally, we employ one hot encoding method to transform human speeches into a term-frequency matrix, whose rows are video ids and columns are words.

Example. To better understand the approach to represent a video as data, consider the example of People.AI. People.AI is a startup founded in 2016 and headquartered in San

Francisco. It applied to Y Combinator acceleration program and successfully got funded. People.AI develops revenue intelligence platform to help its clients to maximize revenue and solve the problem of sales and marketing alignment. Until May 2019, it has raised $97.08M in total.

[Insert Figure 2 and Figure 3 Here.]

We first download People.AI's Y Combinator online application video from YouTube.[7] The video is then decomposed to a series of frames (images) with a sampling rate of 10 frames per second and an audio file. For each frame, we identify all human faces with face detection algorithms. See Figure 2 for the results of face detection for People.AI team using two example frames. We then feed the audio file into the speech-to-text algorithm and obtain a panel for speeches. See Figure 3 for the output of speech-to-text algorithms for People.AI team. Together, we represent the pitch video using three sets of data—faces, audio, and speeches.

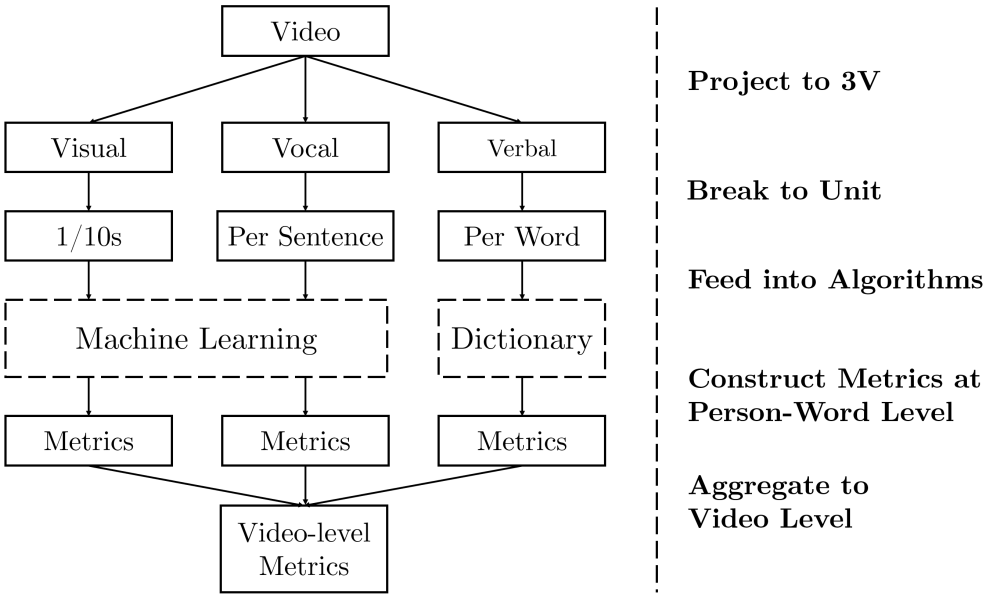
C. Constructing Measures with Machine Learning

In the second step, we construct economically interpretable measures from the represented video data using machine-learning algorithm. One way to think about our machine-learning algorithms is an automatic, tireless, and fair super robot that can rate and record the video data at a high frequency along the three-V dimensions. To do so, we leverage many recent groundbreaking advances in computer vision, speech recognition, and textual analysis.

C.1. Visual. As discussed above, we extract a sequence of images from each video with a sampling rate of ten frames per second. We then identify human faces embedded in each video image. This step is achieved by one class of machine learning algorithms in the field of computer vision, namely face detection. For replicability and transparency, we directly adopt the established implement of Face++. The Face++ platform provides APIs through which we feed our raw images into the cloud computing system and receive a host of face-related measures constructed by Face++'s machine learning algorithms. We also have

[7] Video accessible at https://www.youtube.com/watch?v=61U6VR_6vJU.

Illustration of the Data Processing Framework



robustness checks using an alternative computation platform, Microsoft Azure Cognitive Services.

We obtain the following variables via the Face++ platform. First, via face detection algorithms, the locations of facial landmarks are extracted from the raw images. These coordinates allow us to later detect the movement of eye (e.g., eye blinks) and detect the movement of mouth to match faces with speeches. Second, with emotion recognition algorithms, we are able to characterize facial emotion along the following six dimensions—happiness, sadness, anger, fear, disgust, and neutral. In the empirical analysis, we combine the measures of sadness, anger, fear, and disgust into a composite facial negative emotion measure. We also exclude the neutral emotion in our analysis because it sums up to one along with facial positiveness and negativeness, which induces colinearity. Third, we can also obtain some demographic characteristics of individuals including gender and predicted age.

C.2. Verbal (Text). Next, for verbal, we extract human speeches from audio data by speech-to-text machine learning algorithms. For replicability and transparency, we adopt speech-to-text conversion API of Google Cloud. This machine-learning-powered API converts each audio into a text transcription. These transcripts include a list of words, timestamps (onsets, offsets, and durations) of these words, and punctuation. We then merge the speech

corpus with two dictionaries. The first dictionary is Loughran-McDonald Master Dictionary (LM hereafter), which is commonly used in financial text analysis and provides text categories such as negative, positive, among others (Loughran and McDonald, 2011). The second dictionary is developed by social psychologists using Wordnet and word embeddings (Nicolas, Bai, and Fiske, 2019). This dictionary (NBF hereafter) includes word categories along the dimensions of social psychological traits such as ability and warmth, which helps us to measure verbal characteristics of the social perception (Fiske et al., 2007).

C.3. Vocal (Voice). Finally, we analyze the information embedded in the vocal/audio channel. In this part, we regard audio as digital signals and focus on the information not captured by textual transcription of human speeches. However, different from images which can provide rich information independently, audio data, as sound waves essentially, codes the information in its dynamics and auto-dependence structure. In other words, if we split the audio into fixed high frequency segments, we may lose the information embedded in auto-dependence structure.

To handle this technical problem, we focus on word and sentence unit by leveraging the outputs of speech-to-text algorithms. Specifically, we split each audio into segments by words and sentences of the speech. These units naturally reflect the information structure in the video and are also good approximations of people's cognitive process. That is, when we listen to a speech, we do not hear 10 sampling points of sound waves each second; nor can we attend the entire one-minute sound waves in one second. We process audio signals word by word and sentence by sentence. Therefore, all of our audio analysis is conducted at the word and sentence levels.

We first process audio data and extract useful signals (low-level features). To implement, we employ `pyAudioAnalysis` (Giannakopoulos, 2015), which is a Python package for audio feature extraction, classification, segmentation and applications. Following `pyAudioAnalysis`, we extract 34 low-level features in total. These features include but not limit to spectrogram, chromagram, and energy. These features capture the physical characteristic of the vocal channel.[8]

[8]For a complete list of these audio features, please check `https://github.com/tyiannak/pyAudioAnalysis/wiki/3.-Feature-Extraction`.

We then construct high-level interpretable cognitive measures by feeding these low-level audio features into machine learning algorithms. Specifically, we focus on vocal emotion analysis. We adopt two conceptual frameworks for vocal emotions. The first framework models the vocal emotion along two dimensions—arousal and valance (Bestelmeyer et al., 2017). We use `pyAudioAnalysis`'s established implements and pre-trained machine learning models to obtain vocal arousal and valance. Another framework models the vocal emotion along three emotional dimensions—happy, sad, and neutral. To implement this conceptual model, we employ the implement of `speechemotionrecognition`, which includes deep learning based speech emotion recognition algorithms. We adopt the best pre-trained machine learning models in this package and feed our audio segments into these models. Together, with established machine learning implements, we construct four vocal emotion measures along the dimensions of arousal, valance, happy, and sad.[9] Table A.2 lists the measures constructed in this paper along with their definitions and construction procedures.

D. Measurement Aggregation and Summary Statistics

After constructing measures for each information channel separately, we now merge these measures and aggregate them to the video level. We also consider the cross-channel information structure by using information importance from one channel to weight the measures from other channels in the aggregation.

D.1. Simple Aggregation. We first aggregate all measures into pitch-level using a simple aggregation. The first challenge in doing so is that measures from different information channels have distinct frequencies. In the visual channel, measures are at a frequency of 1/10 second. In audio and text channels, measures are at the level of individual words or sentences. To match the frequencies, we first aggregate image and face measures to the word level according to the onset and offset timestamp of each word. Then the visual data is merged with acoustic data by the timestamp of each word.

It is also worth noting that information from the face channel is person-specific if there are multiple presenters in videos, in which case measures are created for each person in each

[9]The deep learning models in this package are trained on Berlin Database of Emotional Speech (Burkhardt et al., 2005). Vocal neutral is dropped from our analysis due to colinearity.

image. When aggregating face data, we need to take averages across time periods (frames) but within individuals. Together, a video is represented as a panel dataset in our empirical framework. This panel dataset has words as its time indices and human faces as its entity indices. Figure 4 illustrates the data structure and shows how we transform unstructured video data to our final structured panel dataset.

[Insert Figure 4 Here.]

In the empirical analysis, we focus on measures at the pitch-level. We create *Visual-Positive*, *Visual-Negative*, *Vocal-Positive*, and *Vocal-Negative* to capture facial and vocal emotions. For each of these variables, it captures the average proportion of time in the pitch that a team member shows certain facial or vocal emotion by simple averaging the measures at the higher frequency. We create *Visual-Positive* as the score of visual happiness, *Visual-Negative* as the combined score of visual anger, sadness, fear, and disgust. This allows us to mitigate the potential measurement noises created by the machine learning algorithm among subtly different negative visual emotions. For vocal measures, we label Vocal-Happy as *Vocal Positive*, and Vocal-Sad as *Vocal-Negative*. For verbal content, we focus on *Verbal-Positive* and *Verbal-Negative* in the LM financial dictionary and *Verbal-Warmth* and *Verbal-Ability* in the NBF social psychology dictionary. *Verbal-Positive* and *Verbal-Negative* are calculated as the word counts (in each category) scaled by the total number of words in each pitch. *Verbal-Warmth* and *Verbal-Ability* are calculated as the directed word counts (+1, if in the positive direction of the category; -1, if in the negative direction; 0, if unrelated to the category), which are also scaled by the the total number of words. Details of the variable constructions are provided in the Appendix.

Table 3 presents the summary statistics generated for each pitch video. We show, for each feature, the mean, median, 25th and 75th percentiles. The first observation out of this summary statistic table is that there are substantial cross-sectional variations of in startup pitches along the three-V features. Take vocal-happy as an example. The mean 0.140 can be roughly viewed as that in a pitch video, on average the speakers show clear happy vocal features in about 14% of the times. But this number varies quite dramatically. In the lowest extreme case, one speaker could show nearly zero happiness features in voices; while on the

other extreme the happiest team scores 0.396 in the measure.

[Insert Table 3 Here.]

In Table 3 we also show the correlation between metrics from different channels. We find that within a dimension (the same V), features are highly correlated for obvious reasons—videos showing more positive attitude naturally will show less negative attitudes. Meanwhile, across dimensions, similar metrics correlate in very interesting ways. Positive in vocal features is positively correlated with positive visual expressions, confirming the validity of our metrics that are actually generated using totally separate information. But the correlation is fairly mild, suggesting that vocal and facial expressions are separate signals. Textual information is quite independent to vocal and visual features.

D.2. Generating a "Pitch-Factor". Beyond the set of detailed features, we also employ a principal components factor analysis to extract the most important common component from the variance–covariance matrix of the features in pitch videos. This process allows us to detect complex structure in pitches and to eliminate the redundant features. Factor analysis assumes the existence of an underlying pitch factor—a linear combination of the list of three-V features that is not directly observable.

Operationally, factor analysis chooses the vector in the pitch feature space with the greatest variance, where each feature is given equal weight in the total variance calculation. We explore other factor analysis techniques, such as principal factors analysis and maximum-likelihood factor analysis. The qualitative empirical conclusions are not sensitive to the methodology chosen, and the quantitative conclusions change only minimally. For the remainder of this paper, we present the results using the single factor identified by principal components analysis that captures the maximum variation in pitch features. Effectively, principal components analysis performs a singular value decomposition of the correlation matrix of pitch features. The single factor selected in this study is the eigenvector in the decomposition with the highest eigenvalue, we label it as "Pitch-Factor".

The last two columns of Table 3 Panel A reports each variable's loading on the "Pitch-Factor" and the "uniqueness". The factor has a clear and intuitive interpretations. It has

positive loadings on all measures that have positive and affirmative economic meanings. For example, the factor loads positively on *Visual-Positive* (+0.08), *Visual-Beauty* (+0.16), *Vocal-Positive* (+0.37), *Verbal-Warmth* (+0.06) and *Verbal-Ability* (+0.06). Besides, the factor also loads negatively on all measures in the negative direction such as *Visual-Negative* (-0.18), *Vocal-Negative* (-0.29), and *Verbal-Negative* (-0.14). Together, the "Pitch-Factor" can be viewed as a composite index which integrates the information from three-V channels and represents the overall level of positivity and warmth reflected in the pitch video. The uniqueness is the percentage of variance of the variable that cannot be explained by the factor. The uniqueness is low on average so the factor is well-behaving and powerful.

E. Notes on the Methodology

E.1. Connections to the Computer Science Literature. We want to note that this study is highly related to but distinct from the research in computer vision and video processing in the field of computer science. The common thesis for video processing in computer science includes but not limit to object recognition (what objects are in a video), shot transition detection (how many shots are in a video and when are the transitions across these shots), and motion analysis (how to track the movement of objects). These topics are beyond the scope of this paper. Our study is more associated with research in the fields of video indexing, which studies the methods to organize video data and its metadata for high-performance query (Lew, 2013; Hu et al., 2011), and multi-modal analytics and machine learning, which focuses on the integration of information from different perceptual channels (Baltrušaitis et al., 2018).

E.2. Connections to the "Review Rating" Method to Process Video. An emerging literature in psychology and economics uses picture and video data by surveying subjects (reviewers) for ratings. For recent examples, see Graham et al. (2016), Blankespoor et al. (2017), Huang et al. (2018). In these papers, researchers ask subjects to rate speakers based on one frame of video or a short video sample, often along several perception levels such as trustworthiness, likability, attractiveness. Our machine-learning-based methodology has three key differences. First, it allows for ML to scale up the computation and provide a stable and

replicable measurement system. In contrast, the rating method often requires 20-40 subjects to rate one video. Second, it allows for more a fundamental layer of information, along multiple dimensions, rather than only perceptions. Third, the ML-based method mitigates potential biases and inaccuracy from the review and rating process and design (Hall, Mast, and West, 2016).

E.3. Alternative Ways of Introducing Machine Learning. We also want to note the difference between our approach of using ML to construct interpretable economic measure from low-level features (e.g., pixels of images and spectrum of audio frequencies) with an alternative ML approach, which is to directly map them into final outcomes with statistical models (e.g., supervised machine learning). That is, this alternative ML approach can feed our images into a convolutional neural network (CNN) to predict the outcomes (e.g., whether a startup team is invested by the incubator, or whether the company becomes successful).

This approach is straightforward and widely employed in the field of machine learning. Yet it is not ideal in our settings. First, the low-level visual or vocal features used in machine learning algorithms lack support for high-level perceptual labels. Generally, these low-level features (e.g., pixels of images and spectrum of audio frequencies) are hard to interpret per se. And the high dimensionality of these features impedes the usage of statistics and inference methods common in empirical economics studies (e.g., OLS). Second, if we directly link these low-level features to final outcomes under a machine learning framework, we may obtain a statistical model with a good out-of-sample prediction performance. However, these black-box algorithms cannot provide insights about the underlying economic mechanism we are targeting—for example, which feature of faces is most important in explaining final outcomes and what is the economic interpretation of this feature?

5. Empirical Analysis

A. Baseline: Positive Three-V Traits in Pitches Lead to Venture Investment

Our first analysis examines whether features in startup pitches are relevant for investment decisions. We perform the analysis in our full sample using the following specification:

$$I(Invested)_i = \beta_{visual} \cdot X_{visual,i} + \beta_{vocal} \cdot X_{vocal,i} + \beta_{verbal} \cdot X_{verbal,i} + \alpha_{FE} + \varepsilon_i. \quad (1)$$

The key outcome variable of interests is $I(Invested)$, which takes value of 1 if the startup team was chosen by the entrepreneurial investor, and 0 otherwise. Incubator investment typically grants a standard contract of convertible notes with fix amount (usually between $100,000 to $150,000 USD). This shuts down all other potential channels of amount adjustment or term sheet negotiation, making the setting cleaner. On the right-hand side, all pitch features are standardized into a zero-mean variable with a standard deviation of one, which make the economic magnitudes easier to interpret and compare across variables. We estimate Eq. (1) using a logit model, and we focus our attention on the marginal effect of each pitch feature calculated at the sample mean.

We control for fixed effects at the level of incubator to account for the possibility that certain incubators might attract specific types of startup founders or that they have different investment preferences that could correlate with pitch features. Standard errors are clustered at the incubator and year to account for the fact that one applicant's investment decision correlates with others automatically given the quota system in incubator investment.

[Insert Table 4 Here.]

Table 4 presents the results. We focus on visual, vocal, and verbal measures respectively in column (2). A robust pattern emerges from all three-V dimensions: investors are more likely to invest in startups that show more positive attitudes in their pitches. In contrast, startup teams showing more negative attitudes are less likely to be invested.

For the visual feature analysis in column (2), Visual-Positive, capturing happier facial expressions is associated with a higher probability of being invested. The 0.015 coefficient

means that a one standard deviation increase of happiness reflected in the visual dimension is associated with a 1.5 percentage points increase of the investment likelihood. This increase is a 17.6% increase from the unconditional investment probability of 8.52%. On the other end, startup teams that show higher negative facial emotion, as captured using the Visual-Negative, are less likely to be invested. The economic magnitude is even larger, suggesting that the negative component is even more relevant in driving investment decisions.

The effect of visual emotions in pitches are quite sizable. To provide a benchmark, we lean on the well-established concept of "beauty premium" (Hamermesh and Biddle, 1994), which is shown to be important in labor market and other economic decisions. We confirm that more beautiful entrepreneurs are more likely to receive investment. The economic magnitude is roughly the same for Visual-Positive. This means that happy emotion in pitches is economically meaningful in the venture investment decision.

Vocal (Audio) part in column (2) presents the effect of the vocal features. In the vocal dimension, we rely on two vocal emotion categorizations. In the first positive-negative categorization, the pattern is quite similar to the findings about visual features. More positive (negative) tones in pitches are associated with higher (lower) probability of receiving incubator financing. In the second categorization, the audio channel is projected to a two-dimension space of arousal and valence. Valence measures the positive affectivity of the vocal feature, while arousal captures the strength of such an attitude (exciting vs calm)—in some sense, they captures the concept of "passionate" often referenced in entrepreneurship studies (Gompers et al., 2019). We find that high-valence and high-arousal pitches are more likely to attract investment, a one-standard-deviation change in either of the measures is associated with a 2.3 percentage point (27.1% from baseline) increase of the probability of receiving investment.

In the *Verbal (Text)* part of column (2), we focus on another source of rich information—verbal content. The choice of positive vs. negative words in pitches matters for the investment decision. Note that the textual dimensions we explore in this paper are fundamentally different from those that focus on the substance of economic news (Bybee et al., 2019). Instead, we intentionally focus on dimensions that are more relevant for social interactions. The Verbal-Warmth and Verbal-Ability dimensions are closely related to our focus—they are based on the social psychology dictionaries and capture how each word penetrates to

listener's perceptions. Interestingly, warmer pitches (i.e., friendlier and happier) attracts more investment. In contrast, those that choose to talk more about their ability and competitiveness turn away investors more often, which is somewhat surprising given that entrepreneurial investment is a professional decision involving identifying more capable entrepreneurs.

B. Potential Explanations: Startup Quality vs. Social Psychology Forces?

The evidence supports the idea that human interactions are closely linked to economic decisions. The next question is: why? We conduct our exploration by guiding the analysis using two broad and non-mutually-exclusive categories of explanations, which cover several leading possibilities offered by psychology and economics theories.

The first group of explanation concerns the potential information content from human interactions that could help investors better infer the hard-to-observe startup quality. These information could be of various kinds. For example, the attitude of entrepreneurs might reflect the underlying quality of the idea, i.e., entrepreneurs appear to be more positive and more passionate when their startup business if they have private information that the idea is of high type. In addition, the social interactions might reflect the ability and productive personality that are hard to be evaluated without such interactions, such as social skills, positive attitudes, etc. To map to our traditional framework of decision theory and behavioral finance, this is roughly a "beliefs" channel where investors use the pitch to update their beliefs.

The second group of explanations is about the social psychological feature. Specifically, investors, like all human beings, are subject to social psychological forces that may lead them to make investment decisions that are subject to their social psychological limitations but are not necessarily optimal. For instance, people have a natural tendency to find people having "warmer" personalities more likable and trustworthy,

C. Do Pitch Features Capture Omitted Startup Quality?

We first investigate whether the traits in pitches are indeed capturing omitted team quality. If that is the case, the baseline results can be interpreted as that investor use the pitch to better update their beliefs about team quality, which leads to potentially better

investment decisions. Empirically, this can be framed as an omitted variable problem—if there is a perfect signal capturing startup and team quality, the pitches themselves should not matter at all.

The first empirical approach to explore this omitted variable problem is to add control variables that are good proxies of startup and team quality and track the stability of our coefficients of interests. The intuition, which is formalized in a few papers (Altonji, Elder, and Taber, 2005; Oster, 2019) and discussed below, is that if the coefficients are stable to those (good but imperfect) quality controls, that means that the uncovered correlation is not due to the omitted quality variable and is a standalone effect. To implement, we repeat the analysis in columns (1) and (2), with added control variable on founders' education (whether they own an Master or PhD and whether they attend elite universities), founders' entrepreneurship experience, founders' prior work experience. These variables have been shown to be highly correlated with entrepreneurship quality and success.

The results are reported in columns (3) and (4) of Table 4. The first pattern jumping out of the results is that all that we learned from columns (1) and (2) remain robust statistically. But perhaps more importantly, from the perspective of assessing the omitted variable problem, is that the coefficients with and without the controls remain very stable after introducing controls that are well likely correlated with team quality. This evidence, at first glance, suggests that the relation between pitch features and venture investment decision is not hugely affected by the omitted team quality measure.

This argument can be more formally tested based on Oster (2019) and Altonji, Elder, and Taber (2005). The Oster (2019) approach incorporates the change of R^2 induced by adding controls, and argues that the size of R^2 change is informative in judging whether the stability of the estimated coefficient of interest is sufficient to argue away the omitted variable problem. Our test based on the Oster (2019) suggests that the degree of problem is low.

The second test makes use of the long-term performance of startups. If the impact of human interactions on investment decisions is indeed due to the ability to capture a better startup quality, we would expect those valued features to be associated with future success of the company. To be more specific, if feature X positively affect investment decisions because it captures positive startup quality that is not observable, we would expect X to be positively

correlated with ex post performance of the venture.

To perform this exercise, we focus on entrepreneurial ventures in our sample that were invested by a VC. Conditional on being invested, we are interested in the performance of those startups. The analysis is performed using the following model,

$$\begin{aligned} Performance_i = \gamma_{visual} \cdot X_{visual,i} + \gamma_{vocal} \cdot X_{vocal,i} + \gamma_{verbal} \cdot X_{verbal,i} \\ + \gamma \cdot Controls_i + \alpha_{FE} + \varepsilon_i. \end{aligned} \tag{2}$$

The key performance variables for each startup are total employment and total funding raised as of June 2019, and both variables are tracked by PitchBook database and cross-checked with Crunchbase. We use an inverse hyperbolic sine transformation to transform the variables for better empirical properties and the results remain almost identical with alternative transformation methods such as the log transformation. Employment serves as the primary outcome variables. Total financing amount, though often considered as a sign of business success, is less suitable given all the interaction features could potentially affect this variable the same way as in Table 4. On the right hand side, we include all the pitch features, together with startup and team controls used in Table 4. We also include controls of firm age and the squared term of firm age at the time of measurement. We also include industry fixed effects to account for industry-level growth variations at the GICS sector level and accelerator fixed effects.

[Insert Table 5 Here.]

This analysis asks, do firms showing more positivity in pitches indeed have better future performance. Table 5 presents the results and suggests the answer is, *No*. Comparing the coefficients from this performance regression (employment) and the investment regression in Table 4, ten of the total of 11 features flip signs. Some variables turns insignificant due to statistical power issues, while some of them show promising statistical significance. Most noticeably, more positive entrepreneurs, indicated by happier facial expressions and vocal emotions, are in fact less likely to have a growing business. The -0.072 coefficient associated with *Visual-Positive* means that firms with one-standard-deviation more visual positivity at the early stage pitch is 7.2% smaller compared to firms in the same age group and same

industry. In terms of verbal features, those entrepreneurs who appear to be more competitive and highlight their abilities in pitches perform better in growing the business, even though they were less likely to obtain funding. In contrast, warmer pitches lead to higher probability of receiving investment, but they underperform.

Overall, evidence in this section does not seem to lend much support to the explanation that investor invest in startups with more positive pitches due to a better inferred quality. However, this still leaves us two potential possibilities. One, which is also the main focus of the section below, is the impact of social psychology forces that may lead to certain stereotyping, limitations in processing information, among others. The other possibility is that investors think they can infer quality but they do this poorly, i.e., incorrect belief updating from pitches. Disentangling the two is incredible difficult empirically, even with experiment one needs to survey review beliefs directly (Bohren, Haggag, Imas, and Pope, 2019a; Bohren, Imas, and Rosenberg, 2019b). Our analysis going forward is a milder one—we aim to document how some well established social psychological forces interact with the investment decision.

D. Social Psychological Forces

D.1. Gender Stereotype and Differential Effects. We first explore subsamples where social psychological mechanisms may exert impacts differently. If the connection of human interaction features and economic decisions behave differently in those subsamples, it would suggest that social psychology is an important driver of our findings.

We explore this possibility by investigating the role of gender. Social psychology research show that woman are often judged differently and treated differently in social occasions and economic settings such as hiring or promoting. For instance, Fredrickson and Roberts (1997) show that women are often more heavily judged on appearance and non-substantive features, whereas men are not evaluated this way. Women and men are also expected to follow different gender stereotypes—there are generally higher expectations of men in general ability and task performance domains, while women are expected to be high in terms of warmth, empathy, and altruism (Kite, Deaux, and Haines, 2008; Ellemers, 2018).

We first examine whether, as established in social psychology studies, that women are

evaluated differently than men, and whether those patterns coincides with the psychology evidence. In Table 6 we study the heterogeneities of the baseline investment-pitch feature relation across female vs. male entrepreneurs. To perform the analysis, we focus on subsamples of startups with male-only and female-only teams (including one-person startups, but we call these as teams for convenience). For each team, the metrics are naturally calculated for only people of the same gender. We apply the same empirical specification as in Eq. (1). By way of comparing the estimation results for the female and male subsamples, we are able to explore whether certain features are more or less relevant for different gendered entrepreneurs.

[Insert Table 6 Here.]

Table 6 presents the result. We separately report the regression results for both men (columns 1 and 2) and women (columns 3 and 4) entrepreneurs in single-gender teams. In columns (1) and (3) we report how the single factor of pitch features is associated with investment decisions. Given all the variables are already standardized to zero-mean with the standard deviation being one, the difference in the sensitivity is simply reflected on the coefficients. Investment decisions on woman-only startups are in fact more then seven times more sensitive to the performance in the pitch, 0.019 vs. 0.141.

This difference does not only hold in the factor analysis. In fact, it holds in a feature-by-feature analysis, as shown in column (2) and (4). For example, investment decisions are more sensitive to the level of facial positivity among female entrepreneurs than on male entrepreneurs. In other words, if female entrepreneurs do not appear to be positive in their pitches, the negative consequence, i.e., a lower probability of obtaining venture financing, is much higher compared to male investors. Some of the estimated coefficients are statistically weak due to power issues, but are economically sizable. We do observe, however, some coefficients associated with features of man-only teams no longer matter economically. For instance, the positivity in vocal and verbal matter little in these team, while they do

This table is consistent with the literature on gender stereotyping, because of which women and men are evaluated differently in social interactions and economic decisions. Not only that women are more judged based on appearance and non-substantive feature in pitches, but also that the sensitivity is in the same direction as the gender stereotype—investors reward

women who confirms their stereotypes, i.e., more warmth, more positive, among others.

What if a team has both male and female entrepreneurs? In Table 7 we focus our study on the subsample of entrepreneurial teams that have both male and female team members pitching. For each team, we separately calculate the three-V dimensions for women in the team and men in the team. We put those measures jointly in Eq. (1) so we can see that whether the features from women or men carry the important weight for the probability that the team being invested.

[Insert Table 7 Here.]

Table 7 shows the results. At the first glance, comparing columns (1) and (3), men drive the majority of the relations between pitch features and investment decisions. In other words, women are more neglected in social interactions when co-presenting with a man. Note that this is after the fact that men and women actually speak for similar amount of time in pitches on average. In columns (2) and (4), we see this neglect mainly comes from the visual channel, and investors are more likely to invest in male entrepreneurs show positive attitudes; while at the same time, the coefficient associated with This is perhaps not surprising to people familiar with entrepreneurship anecdotes, but this is, to our knowledge, the first evidence to show this neglect of women.

When examining the vocal domain, the coefficients associated with *Vocal-Arousal* and *Vocal-Valence* carry opposite and significant coefficients compared to the baseline regressions in Table 4 and in Table 6. That is, investors invest in teams with less passionate and less energetic woman entrepreneurs when presenting with men. One interpretation is related to team dynamics—investors prefer that, in a mixed-gender team, men are more powerful, passionate, and even charismatic, while women are in a secondary, supportive or even obedient role. In fact, if women appear to be a powerful leader, that lowers the probability of receiving funding.

Overall, the evidence suggests that gender stereotyping and inequality play a role in governing how pitches affect investment decisions. One further test of this mechanism is to examine the role of reviewers' gender. This is not available from other setting since the decision of investment is often made by groups rather than one individual. What is reassuring,

however, is that the gender-related social psychology forces are often robust among male and female reviewers (Hentschel, Heilman, and Peus, 2019).

D.2. First Impression and Superficial Impression. Our next set of test tries to explore another source of variation built in the process that humans process information and form perceptions. The idea is, if pitches are used as a source of information on the quality of the entrepreneurial team or the idea itself, one would expect that the deeper discussion on the business ideas, and the pitch performance therein, should carry more weight in predicting investment decisions. In contrast, if investors simply use pitches as a way to form an impression without intention of analyzing information, we would expect the less accurate signals to perform as well in predicting investment behaviors.

We next study the role of forming first impressions. In social psychology, there is a large literature documenting the fast formation of first impressions, which in turn affect perceptions and decisions. This first impression sometimes are not based on substantive information, but rather through "superficial" features such as the facial expression or the way of greeting. In economic literature, Rabin and Schrag (1999) show that even when first impression lacks accuracy or even wrong, confirmation bias could make the impact of long-lasting.

In our setting, we are curious to see if the investors' decisions are responsive to just the first five seconds of the pitches. In those five seconds, pitches include only the greetings part of the content and at most a brief introduction of the team. In other words, the informational content of these five seconds is low. We recalculate the measures using only the first five seconds of the video, and use them in the same model Eq. (1) and report the results in Table 8 columns (3) and (4). For comparison, columns (1) and (2) reproduce the baseline results in Table 4.

[Insert Table 8 Here.]

The result based on the Pitch-Factor suggest that the first five seconds of the pitch already are enough for investors to make up their decisions. The economic magnitude are close and the result is statistically robust. The effect of the first five seconds concentrates in mostly visual, but also vocal features. It shows up noisily in textual features. This is consistent

with the social psychology literature that facial expressions are the primary determinants of first impressions (Todorov et al., 2005; Willis and Todorov, 2006; Todorov and Porter, 2014). We show that the correlation between the first five seconds of video and the full video is only mild (0.5-0.6). Overall, our evidence is consistent with that investors quickly form perceptions from the first few seconds, rather than examining the full videos.

We next directly test whether investors pay extra attention to informative materials. The idea is, if investors intend to extract signals of private information of entrepreneurs about the company, they would value more heavily the signal from the information-intense or business-related parts (say introducing the business idea, etc.) compared to the signals from relatively low-information parts (the superficial impressions, say greetings, introducing team members, etc.).

To perform this test, one need to separately construct measures for periods when the startup teams talk about more and less useful topics. We leverage the machine learning algorithm's ability to simultaneously record the content and pitch features, and construct a content-weighted measure. If investor indeed extract entrepreneurs' private information from the interaction, we would expect the content-weighted feature to be more related to investment decisions than the generic measures used in Table 4. To construct the content-weighted measures, we calculate term frequency-inverse document frequency (*TF-IDF*). We use the speech corpus extracted from the pitching videos in our sample. The intuition is that a word would be more important in a specific video if it appears many times in this video and does not show up a lot in other videos.

Table 8 columns (5) and (6) present the result. If investors are indeed attempting to extract information about the quality of the business or the ability of the entrepreneurs, one would expect the TF-IDF version, which over-weigh business-related information, be more powerful in explaining the results. Our analysis does not support this conjecture.

6. Discussion and Conclusion

To conclude, we provide some thoughts on how this work may fit a broader trend of studying social psychology forces in economics and finance, and how the empirical methodology

can be extended to future work.

A. Connections to Social Psychology

Conceptually, this paper makes a connection between social psychology and economic decisions in the real world. Social psychology studies how people's thoughts, feelings, and behaviors are influenced by the actual, imagined or implied presence of others. Even though lab experiments help establish many important principles, how those mechanisms affect real world remain understudied. The behavioral finance literature has mainly focused on cognitive biases and their impact on financial investment decisions. Social psychology has largely been overlooked in such settings, especially empirically. This is largely due to the lack of data and empirical design. This paper connects to those streams of literature using the novel empirical framework.

One can imagine the development of the connection between social psychology and economics and finance research to benefit from the theories and empirical methodologies established by the behavioral economics. For example, to extend our understanding on the effect of social interactions, we should better study whether those communication features are part of investors' preferences or formation of beliefs. Answering these questions may require a combination of the video-based approach with modeling or experiments.

B. Methodological Extension to More Complex Interaction Settings

Human interactions are complex activities. In a general setting, human interactions involve multiple players, who sequentially send and receive signals via the three dimensions, which affect each others' decisions. The setting studied here is one of the simplest, in which one player (the entrepreneur) sends out a signal to the other end of the player (the investor). We are examining the communication made from this one end to the other.

The extension of the framework to a more complex setting is natural. Imagine that one could obtain videos for human interactions with two players, the processing could be performed on different players separately and then aggregated for them respectively. One could imagine this process be done for the court, for negotiations, for classrooms, for interviews, among others.

C. Suggestions for Future Venture Investment Practice

At last, as the venture investment field involves large uncertainty and lack of information, investment decisions often are based on noisy signals. This leaves room for social psychology-driven decisions, which may lead to biases against certain population such as female entrepreneurial or minority entrepreneurs. This may also lead to worse investment performance from the investor side.

One natural implication of our paper is to manage the use of pitches in venture financing, especially in earlier stage. Potential changes could be introducing multiple reviewers with different demographics and cultural backgrounds, the reduces use of certain channels in pitches (such as limiting the communication to phone call only), among others.

References

Adams, Reginald B, Reginald B Adams Jr, Nalini Ambady, Shinsuke Shimojo, and Ken Nakayama, 2011, *The Science of Social Vision: The Science of Social Vision*, volume 7 (Oxford University Press).

Altonji, Joseph G, Todd E Elder, and Christopher R Taber, 2005, An evaluation of instrumental variable strategies for estimating the effects of catholic schooling, *Journal of Human resources* 40, 791–821.

Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency, 2018, Multimodal machine learning: A survey and taxonomy, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 423–443.

Bernstein, Shai, Arthur Korteweg, and Kevin Laws, 2017, Attracting early-stage investors: Evidence from a randomized field experiment, *The Journal of Finance* 72, 509–538.

Bestelmeyer, Patricia EG, Sonja A Kotz, and Pascal Belin, 2017, Effects of emotional valence and arousal on the voice perception network, *Social cognitive and affective neuroscience* 12, 1351–1358.

Blankespoor, Elizabeth, Bradley E Hendricks, and Gregory S Miller, 2017, Perceptions and price: Evidence from ceo presentations at ipo roadshows, *Journal of Accounting Research* 55, 275–327.

Bohren, J Aislinn, Kareem Haggag, Alex Imas, and Devin G Pope, 2019a, Inaccurate statistical discrimination, *NBER Working Paper* .

Bohren, J Aislinn, Alex Imas, and Michael Rosenberg, 2019b, The dynamics of discrimination: Theory and evidence, *American Economic Review* 109, 3395–3436.

Brooks, Alison Wood, Laura Huang, Sarah Wood Kearney, and Fiona E Murray, 2014, Investors prefer entrepreneurial ventures pitched by attractive men, *Proceedings of the National Academy of Sciences* 111, 4427–4431.

Burkhardt, Felix, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss, 2005, A database of german emotional speech, in *Ninth European Conference on Speech Communication and Technology*.

Bybee, Leland, Bryan T Kelly, Asaf Manela, and Dacheng Xiu, 2019, The structure of economic news, *Available at SSRN 3446225* .

Cuddy, Amy JC, Peter Glick, and Anna Beninger, 2011, The dynamics of warmth and competence judgments, and their outcomes in organizations, *Research in organizational behavior* 31, 73–98.

Ellemers, Naomi, 2018, Gender stereotypes, *Annual review of psychology* 69, 275–298.

Fiske, Susan T, Amy JC Cuddy, and Peter Glick, 2007, Universal dimensions of social cognition: Warmth and competence, *Trends in cognitive sciences* 11, 77–83.

Fredrickson, Barbara L, and Tomi-Ann Roberts, 1997, Objectification theory: Toward understanding women's lived experiences and mental health risks, *Psychology of women quarterly* 21, 173–206.

Gandomi, Amir, and Murtaza Haider, 2015, Beyond the hype: Big data concepts, methods, and analytics, *International journal of information management* 35, 137–144.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, 2019, Text as data, *Journal of Economic Literature* 57, 535–74.

Giannakopoulos, Theodoros, 2015, pyaudioanalysis: An open-source python library for audio signal analysis, *PloS one* 10, e0144610.

Gompers, Paul A, Will Gornall, Steven N Kaplan, and Ilya A Strebulaev, 2019, How do venture capitalists make decisions?, *Journal of Financial Economics* .

Graham, John R, Campbell R Harvey, and Manju Puri, 2016, A corporate beauty contest, *Management Science* 63, 3044–3056.

Hall, Judith A, Marianne Schmid Mast, and Tessa V West, 2016, *The social psychology of perceiving others accurately* (Cambridge University Press).

Hamermesh, Daniel S, and Jeff E Biddle, 1994, Beauty and the labor market, *American Economic Review* 1174–1194.

Hentschel, Tanja, Madeline E Heilman, and Claudia V Peus, 2019, The multiple dimensions of gender stereotypes: a current look at men's and women's characterizations of others and themselves, *Frontiers in psychology* 10, 11.

Hu, Weiming, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank, 2011, A survey on visual content-based video indexing and retrieval, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 41, 797–819.

Huang, Laura, and Jone L Pearce, 2015, Managing the unknowable: The effectiveness of early-stage investor gut feel in entrepreneurial investment decisions, *Administrative Science Quarterly* 60, 634–670.

Huang, Xing, Zoran Ivković, J Jiang, and I Wang, 2018, Swimming with the sharks: Entrepreneurial investing decisions and first impression .

Kanze, Dana, Laura Huang, Mark A Conley, and E Tory Higgins, 2018, We ask men to win and women not to lose: Closing the gender gap in startup funding, *Academy of Management Journal* 61, 586–614.

Kaplan, Steven N, Mark M Klebanov, and Morten Sorensen, 2012, Which ceo characteristics and abilities matter?, *The Journal of Finance* 67, 973–1007.

Kaplan, Steven N, and Per Stromberg, 2001, Venture capitals as principals: contracting, screening, and monitoring, *American Economic Review* 91, 426–430.

Kaplan, Steven N, and Per ER Strömberg, 2004, Characteristics, contracts, and actions: Evidence from venture capitalist analyses, *Journal of Finance* 59, 2177–2210.

Kite, Mary E, Kay Deaux, and Elizabeth L Haines, 2008, Gender stereotypes, *Psychology of women: A handbook of issues and theories* 2, 205–236.

Knapp, Mark L, Judith A Hall, and Terrence G Horgan, 2013, *Nonverbal communication in human interaction* (Cengage Learning).

Krauss, Robert M, William Apple, Nancy Morency, Charlotte Wenzel, and Ward Winton, 1981, Verbal, vocal, and visible factors in judgments of another's affect., *Journal of Personality and Social Psychology* 40, 312.

Lew, Michael S, 2013, *Principles of visual information retrieval* (Springer Science & Business Media).

Loughran, Tim, and Bill McDonald, 2011, When is a liability not a liability? textual analysis, dictionaries, and 10-ks, *The Journal of Finance* 66, 35–65.

Malmendier, Ulrike, and Geoffrey Tate, 2005, Ceo overconfidence and corporate investment, *Journal of Finance* 60, 2661–2700.

Malmendier, Ulrike, and Geoffrey Tate, 2008, Who makes acquisitions? ceo overconfidence and the market's reaction, *Journal of Financial Economics* 89, 20–43.

Malmendier, Ulrike, and Geoffrey Tate, 2015, Behavioral ceos: The role of managerial overconfidence, *Journal of Economic Perspectives* 29, 37–60.

Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers, 2011, Big data: The next frontier for innovation, competition, *Washington, DC: McKinsey Global Institute* .

Mayew, William J, and Mohan Venkatachalam, 2012, The power of voice: Managerial affective states and future firm performance, *Journal of Finance* 67, 1–43.

McArthur, Leslie Z, and Reuben M Baron, 1983, Toward an ecological theory of social perception, *Psychological Review* 90, 215.

Mehrabian, Albert, 1972, *Nonverbal Communication* (Transaction Publishers).

Mobius, Markus M, and Tanya S Rosenblat, 2006, Why beauty matters, *American Economic Review* 96, 222–235.

Murphy, Nora A, Judith A Hall, and C Randall Colvin, 2003, Accurate intelligence assessments in social interactions: Mediators and gender effects, *Journal of Personality* 71, 465–493.

Nicolas, Gandalf, Xuechunzi Bai, and Susan T Fiske, 2019, Automated dictionary creation for analyzing text: An illustration from stereotype content .

Oster, Emily, 2019, Unobservable selection and coefficient stability: Theory and evidence, *Journal of Business & Economic Statistics* 37, 187–204.

Rabin, Matthew, and Joel L Schrag, 1999, First impressions matter: A model of confirmatory bias, *The quarterly journal of economics* 114, 37–82.

Strahan, Carole, and Donald G Zytowski, 1976, Impact of visual, vocal, and lexical cues on judgments of counselor qualities., *Journal of Counseling Psychology* 23, 387.

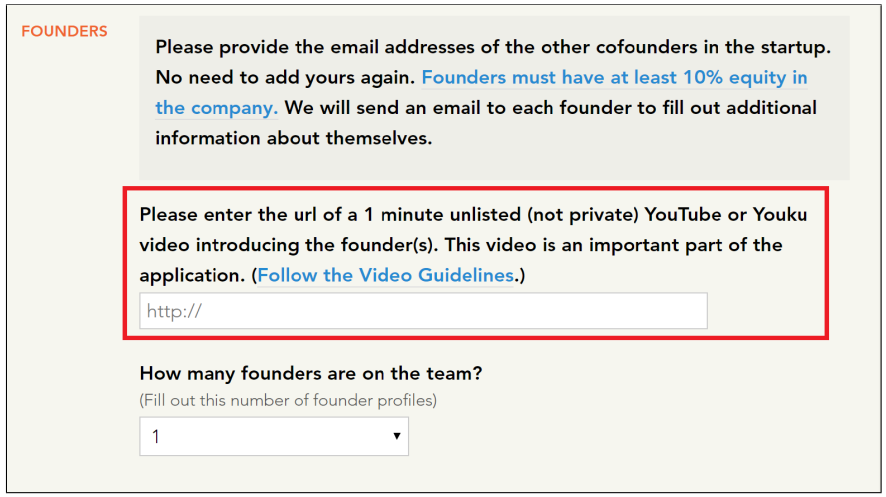
Todorov, Alexander, Anesu N Mandisodza, Amir Goren, and Crystal C Hall, 2005, Inferences of competence from faces predict election outcomes, *Science* 308, 1623–1626.

Todorov, Alexander, and Jenny M Porter, 2014, Misleading first impressions: Different for different facial images of the same person, *Psychological science* 25, 1404–1417.

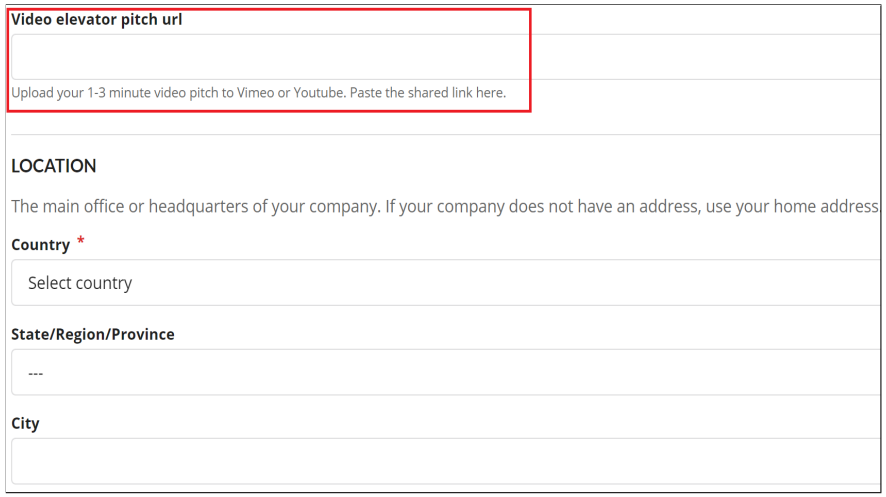
Willis, Janine, and Alexander Todorov, 2006, First impressions: Making up your mind after a 100-ms exposure to a face, *Psychological science* 17, 592–598.

Figure 1. Examples of Incubator Online Application Forms

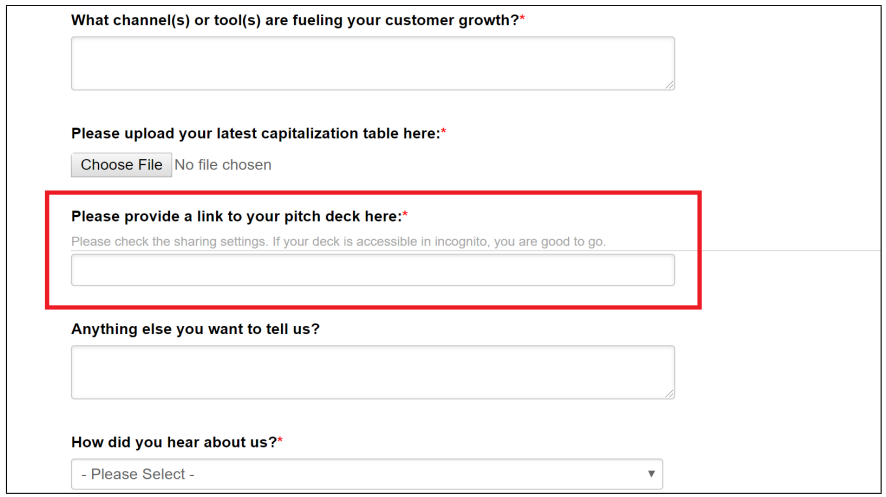
(a) Y Combinator



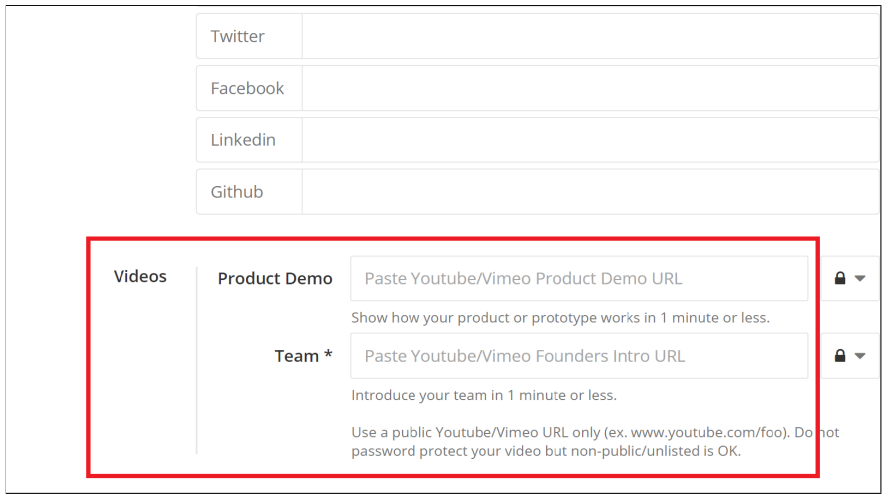
(b) MassChallenge



(c) 500 Startups

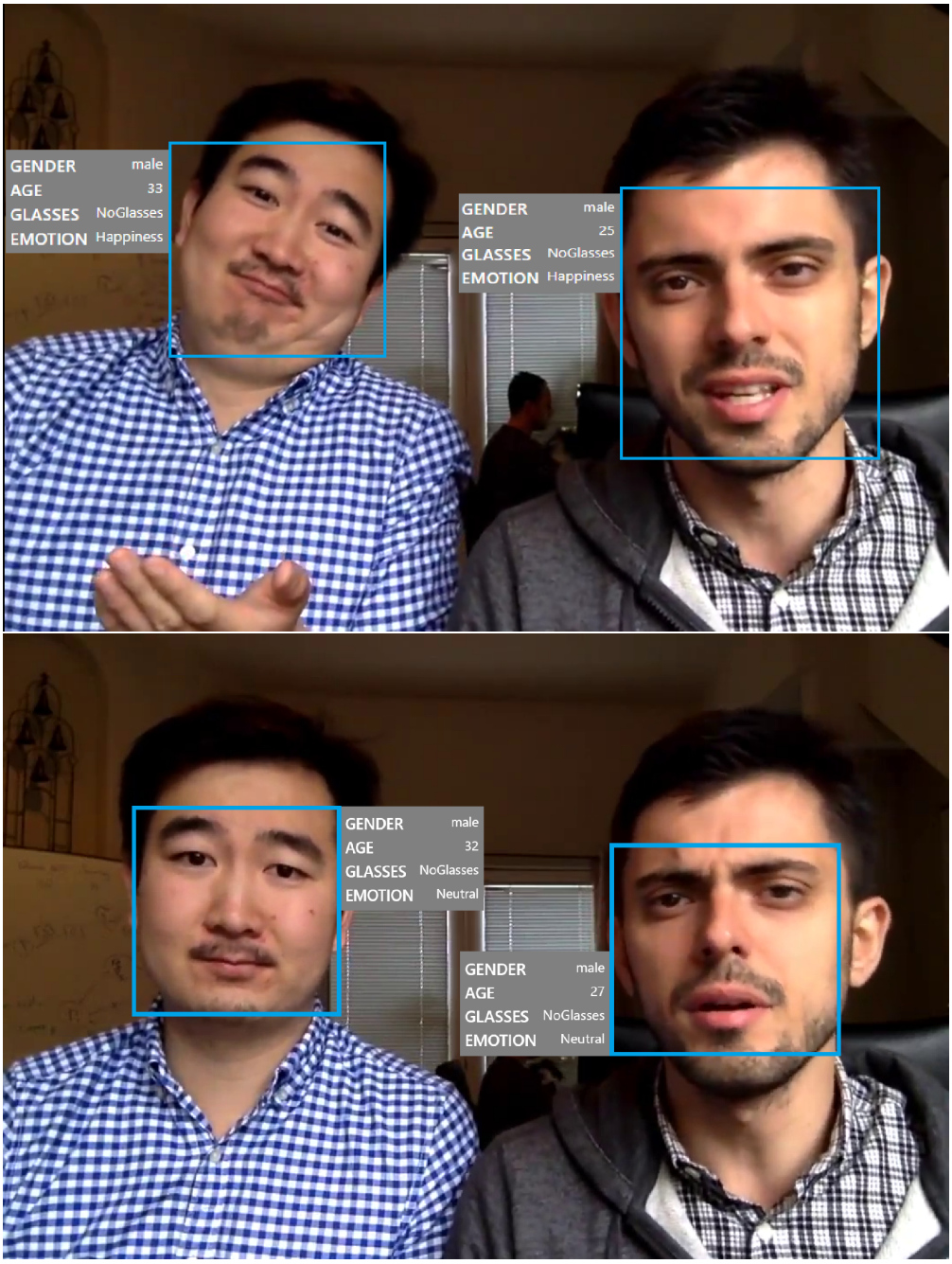


(d) Techstars



Notes: This figure shows screenshots of incubators' online application forms. We show forms for Y Combinator, MassChallenge, 500 Startups, and Techstars. On each online application form, we highlight the question specifically asking for uploadig pitch videos.

Figure 2. Examples of Face Detection



Notes: This figure presents examples of the output from the face detection algorithm. The algorithm detects human faces in images along with face attribute features such as gender, age, and facial emotion.

Figure 3. Examples of Speech-to-text Output

Text	Onset	Offset	Duration	Confidence	Sentence End	Speaker ID
Hi	0.3	0.7	0.4	0.91	YES	1
I'm	0.7	0.9	0.2	0.91	NO	1
Oleg	0.9	1.5	0.6	0.91	NO	1
co-founder	1.5	2	0.5	0.91	NO	1
and	2	2.1	0.1	0.91	NO	1
CEO	2.1	2.6	0.5	0.91	NO	1
of	2.6	2.8	0.2	0.91	NO	1
people	2.8	3.1	0.3	0.91	NO	1
dot	3.1	3.2	0.1	0.91	NO	1
AI	3.2	3.5	0.3	0.91	YES	1
I	3.5	3.8	0.3	0.91	NO	1
have	4	4.2	0.2	0.91	NO	1
ten	4.2	4.4	0.2	0.91	NO	1
years	4.4	4.8	0.4	0.91	NO	1
of	4.8	4.9	0.1	0.91	NO	1
enterprise	4.9	5.3	0.4	0.91	NO	1
sales	5.3	5.6	0.3	0.91	NO	1
experience	5.6	6.1	0.5	0.91	NO	1
selling	6.1	6.4	0.3	0.91	NO	1
machine	6.4	6.7	0.3	0.91	NO	1
learning	6.7	7.1	0.4	0.91	NO	1
...	...	...	...	...	...	...
Hi	11.2	11.5	0.3	0.91	YES	2
I'm	11.5	11.7	0.2	0.91	NO	2
Kevin	11.7	12	0.3	0.91	NO	2
co-founder	12	12.4	0.4	0.91	NO	2
and	12.4	12.5	0.1	0.91	NO	2
CTO	12.5	13.1	0.6	0.91	YES	2
I	13.3	13.4	0.1	0.91	NO	2
built	13.4	13.5	0.1	0.91	NO	2
a	13.5	13.6	0.1	0.91	NO	2
recommendation	13.6	14.2	0.6	0.91	NO	2
engine	14.2	14.5	0.3	0.91	NO	2
in	14.5	14.6	0.1	0.91	NO	2

Text	Onset	Offset	Duration	Confidence	Sentence End	Speaker ID
two	14.6	14.8	0.2	0.91	NO	2
weeks	14.8	15	0.2	0.91	NO	2
that	15	15.2	0.2	0.91	NO	2
beat	15.2	15.3	0.1	0.91	NO	2
the	15.3	15.4	0.1	0.91	NO	2
YouTube	15.4	15.8	0.4	0.91	NO	2
and	15.8	16	0.2	0.91	NO	2
Netflix	16	16.4	0.4	0.91	NO	2
teams	16.4	16.9	0.5	0.91	YES	2
...	...	...	...	...	...	...
there	26.4	26.5	0.1	0.91	NO	2
is	26.5	26.7	0.2	0.91	NO	2
a	26.7	26.7	0	0.91	NO	2
fundamental	26.7	27.3	0.6	0.91	NO	2
flaw	27.3	27.9	0.6	0.91	NO	2
to	27.9	28	0.1	0.91	NO	2
having	28	28.3	0.3	0.91	NO	2
professionally	28.3	28.9	0.6	0.91	NO	2
trained	28.9	29.3	0.4	0.91	NO	2
sellers	29.3	30	0.7	0.91	NO	2
sell	30	30.3	0.3	0.91	NO	2
themselves	30.3	31	0.7	0.91	YES	2
...	...	...	...	...	...	...
We	33.3	33.4	0.1	0.91	NO	1
built	33.4	33.6	0.2	0.91	NO	1
a	33.6	33.6	0	0.91	NO	1
platform	33.6	34.1	0.5	0.91	NO	1
that	34.1	34.2	0.1	0.91	NO	1
can	34.2	34.3	0.1	0.91	NO	1
quantify	34.3	34.7	0.4	0.91	NO	1
salesperson	34.7	35.3	0.6	0.91	NO	1
skills	35.3	35.8	0.5	0.91	NO	1
and	36	36.1	0.1	0.91	NO	1
networks	36.1	36.7	0.6	0.91	NO	1

Notes: This figure presents an example of the output from the speech-to-text algorithm. The algorithm identifies the words spoken, the time stamps (the onset and offset of each word and the duration), punctuation, and speaker id.

Figure 4. Data Structure and Processing Procedure

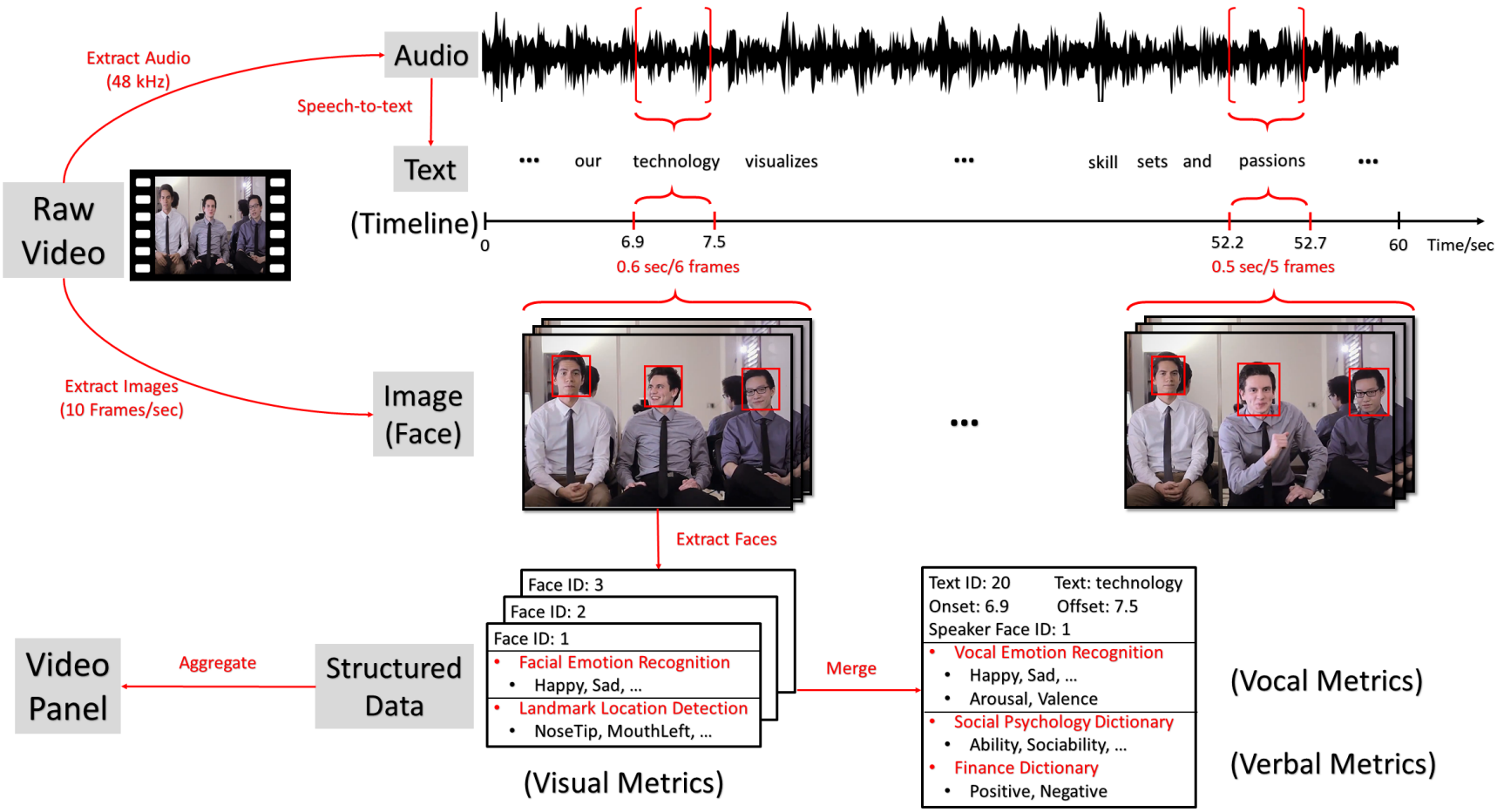


Table 1. Sample Description of Pitch Videos

Panel A: Breakdown by Accelerators and Investment Status

Accelerator	Videos #	Accelerator Invested	Website Active	In Crunchbase	In Pitchbook
500 Startups	33	1	15	19	8
AngelPad	83	2	33	36	18
MassChallenge	166	56	129	113	79
Techstars	136	3	67	53	21
Y Combinator	713	35	363	238	91
YC Fellowship	8	0	2	3	0
Total	1,139	97	609	462	217
% of Full Sample	100.00%	8.52%	53.47%	40.56%	19.05%

Panel B: Breakdown by Years

Accelerator	<=2012	2013	2014	2015	2016	2017	2018	2019
500 Startups	1	1	7	7	2	8	5	2
AngelPad	11	7	13	4	12	14	21	1
MassChallenge	4	9	4	13	34	33	34	35
Techstars	9	17	12	15	8	30	32	13
Y Combinator	10	31	29	82	67	110	164	220
YC Fellowship	0	0	0	8	0	0	0	0
Total	35	65	65	129	123	195	256	271
% of Full Sample	3.07%	5.71%	5.71%	11.33%	10.80%	17.12%	22.48%	23.79%

Notes: This table provides descriptive statistics on collected videos by incubators that the applications are made to (Panel A) and by year (Panel B). We obtain pitch videos using an automatic searching script for two public video-sharing websites, YouTube and Vimeo. Integrated with query APIs, our web crawler returns a list of video indices according to a lot of pre-defined keywords, which include but not limit to the names of these accelerators, "startup accelerator application video","accelerator application videos" and so on. We first obtain the full list of potential videos returned by each keyword search (there is a limitation of returned videos by YouTube), and then filter the potential videos by a combination of different conditions on video info obtained along with the video itself. Filtering variables include but not limit to data format, duration, title, and annotation. It should be noted that because Y Combinator has too many application videos each year, we specifically specify the years in the searching keywords so as to obtain more complete results. We also obtain additional videos from incubators' website. Panel A reports the number of videos made to each accelerator and the proportion of each accelarator in the full sample. Panel B reports the breakdown by the application year (typically the year of video uploading).

Table 2. Summary Statistics of Videos and Start-ups

Panel A: Summary Statistics of Video Pitches

	N	Mean	STD	25%	50%	75%
Duration (second)	1,139	83.43	39.51	60.00	68.00	97.00
Video Size (MB)	1,139	18.16	18.37	6.88	12.86	22.64
Frames (per second)	1,139	28.36	7.63	29.00	29.97	30.00
Number of People	1,139	1.74	0.84	1.00	2.00	2.00
Number of Words	1,139	228.53	107.76	163.00	201.00	262.00
Number of Sentences	1,139	15.91	7.33	11.00	14.00	19.00
Number of Views (YouTube)	1,139	764.37	6956.02	31.00	79.00	197.00
Number of Likes (YouTube)	1,139	1.51	6.60	0.00	0.00	1.00
Number of Dislikes (YouTube)	1,139	0.15	0.65	0.00	0.00	0.00

Panel B: Summary Statistics of Start-ups

	N	Mean	STD	25%	50%	75%
Firm Age	1,139	3.20	2.01	2.00	3.00	5.00
Total Funding Amount (K)	208	11,804	66,794	108	382	2,125
Total Funding Rounds	293	2.00	1.40	1	1	2
Number of Employees	424	21.16	70.48	5	5	30

Panel C: Summary Statistics of Teams

	N	Mean	STD	25%	50%	75%
Single-Member	1,139	0.46	0.50	0.00	0.00	1.00
Multi-Member	1,139	0.54	0.50	0.00	1.00	1.00
Men-Only	1,139	0.49	0.50	0.00	0.00	1.00
Women-Only	1,139	0.27	0.45	0.00	0.00	1.00
Mixed Gender	1,139	0.24	0.43	0.00	0.00	0.00
White-Only	1,139	0.50	0.50	0.00	0.00	1.00
Non-White-Only	1,139	0.37	0.48	0.00	0.00	1.00
Mixed Race	1,139	0.13	0.34	0.00	0.00	0.00
Prior Senior Position	1,139	0.47	0.50	0.00	0.00	1.00
Prior Startup Experience	1,139	0.30	0.46	0.00	0.00	1.00
Elite University	1,139	0.06	0.24	0.00	0.00	0.00
Master Degree	1,139	0.19	0.40	0.00	0.00	0.00
PhD Degree	1,139	0.03	0.17	0.00	0.00	0.00

Notes: This table provides descriptive statistics of pitch videos and the underlying startups in our sample. For each variable, we report the number of observations, mean, standard deviation, 25th, 50th, and 75th percentiles. In Panel A, we report the basic information of the pitch videos, including the duration, number of views, number of likes and dislikes. We also report basic information about the pitch, including the number of words and sentences, as well as the team composition categorized as the number of team members and gender ratio. In Panel B, we report the characteristics of startups, including Total Funding Amount, Total Funding Rounds, and Number of Employees, all measured as of 2019. These information is collected from Crunchbase and Pitchbook. In Panel C we report the summary statistics of the startup teams.

Table 3. Summary Statistics of Pitching Behavior Metrics

Panel A: Summary Statistics of Unstandardized Features

	N	Mean	STD	25%	50%	75%	Pitch-Factor Loading	Uniqueness
Visual (Facial)								
Visual-Positive	1,139	0.17	0.16	0.05	0.12	0.25	0.08	0.25
Visual-Negative	1,139	0.15	0.14	0.06	0.11	0.20	-0.18	0.35
Visual-Beauty	1,139	0.58	0.08	0.54	0.59	0.64	0.16	0.45
Vocal (Audio)								
Vocal-Positive	1,139	0.09	0.05	0.06	0.08	0.12	0.37	0.43
Vocal-Negative	1,139	0.01	0.01	0.01	0.01	0.02	-0.29	0.41
Vocal-Arousal	1,139	0.55	0.27	0.39	0.58	0.76	0.91	0.15
Vocal-Valence	1,139	0.44	0.22	0.31	0.46	0.59	0.88	0.16
Verbal (Text)								
Verbal-Positive	1,139	0.01	0.01	0.01	0.01	0.02	0.02	0.33
Verbal-Negative	1,139	0.01	0.01	0.00	0.01	0.01	-0.14	0.40
Verbal-Warmth	1,139	0.02	0.01	0.01	0.01	0.02	0.06	0.31
Verbal-Ability	1,139	0.03	0.02	0.02	0.03	0.05	0.06	0.28

Panel B: Correlations of the Features

	(1)	(2)	(3)	(4)	(5)
(1) Visual-Positive	1.00				
(2) Visual-Negative	-0.12***	1.00			
(3) Visual-Beauty	-0.02	-0.20***	1.00		
(4) Vocal-Positive	0.16***	0.07**	-0.05*	1.00	
(5) Vocal-Negative	0.05*	0.06**	0.01	-0.07**	1.00
(6) Vocal-Arousal	0.02	-0.07**	0.05*	0.24***	-0.15***
(7) Vocal-Valence	-0.02	-0.07**	0.09***	0.13***	-0.12***
(8) Verbal-Positive	0.01	0.03	-0.01	0.02	-0.06*
(9) Verbal-Negative	-0.10***	0.04	-0.01	0.02	-0.04
(10) Verbal-Warmth	-0.05	-0.00	0.01	-0.01	-0.01
(11) Verbal-Ability	0.00	0.02	-0.02	0.04	0.02
Continued	(6)	(7)	(8)	(9)	(10)
(6) Vocal-Arousal	1.00				
(7) Vocal-Valence	0.75***	1.00			
(8) Verbal-Positive	-0.01	0.01	1.00		
(9) Verbal-Negative	-0.08***	-0.07**	0.00	1.00	
(10) Verbal-Warmth	0.02	0.04	0.03	-0.05*	1.00
(11) Verbal-Ability	0.01	0.04	0.08***	-0.03	-0.02

Notes: This table provides summary statistics of the human interaction features. In Panel A, for each variable, we report the number of observations, mean, standard deviation, 25th, 50th, and 75th percentiles. Variables are categorized into vocal, video, and verbal. Panel B provides correlations of the human interaction features listed in Table A.2. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4. Features in Pitches and Investment Decisions

	(1)	(2)	(3)	(4)
	Logistic Marginal Effect. Dependent Var: *I(Invested)*			
Pitch-Factor	0.030***		0.026***	
	(0.007)		(0.007)	
Visual (Facial)				
Visual-Positive		0.015***		0.013**
		(0.005)		(0.006)
Visual-Negative		-0.027***		-0.029***
		(0.007)		(0.007)
Visual-Beauty		0.015**		0.015**
		(0.006)		(0.006)
Vocal (Audio)				
Vocal-Positive		0.009**		0.011*
		(0.005)		(0.006)
Vocal-Negative		-0.045***		-0.047***
		(0.016)		(0.017)
Vocal-Arousal		0.023***		0.019**
		(0.009)		(0.008)
Vocal-Valence		0.023***		0.020***
		(0.006)		(0.007)
Verbal (Text)				
Verbal-Positive		-0.010		-0.011
		(0.009)		(0.009)
Verbal-Negative		-0.026***		-0.023***
		(0.007)		(0.007)
Verbal-Warmth		0.026***		0.027***
		(0.008)		(0.008)
Verbal-Ability		-0.049***		-0.043***
		(0.009)		(0.007)
Observations	1,139	1,139	1,139	1,139
Video Characteristics			Y	Y
Firm/Team Controls			Y	Y
Accelerator FE	Y	Y	Y	Y

Notes: Logistic regressions, marginal effect. The analysis is obtained using the following model:

$$I(Invested) = \beta_{vocal} \cdot X_{vocal} + \beta_{vision} \cdot X_{vision} + \beta_{text} \cdot X_{text} + \alpha_{FE} + \varepsilon.$$

$I(Invested)$ takes value of one if the startup team was chosen by the accelerator, and 0 otherwise. All pitch feature variables are standardized into a zero-mean variable with a standard deviation of 1. Standard errors clustered at the accelerator and year level are displayed in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

Table 5. Long-term Performance of Startups and Features in Pitches

	Flip Sign w/ *I(Invested)*?	(1)	(2)	(3)	(4)
		Dep Var: *Employment*		Dep Var: *Total Financing Amt*	
Pitch-Factor	Y	-0.192**		-0.043	
		(0.056)		(0.103)	
Visual (Facial)					
Visual-Positive	Y		-0.072*		-0.134**
			(0.033)		(0.044)
Visual-Negative	Y		0.141		0.015
			(0.096)		(0.070)
Visual-Beauty	Y		-0.059		0.156
			(0.133)		(0.130)
Vocal (Audio)					
Vocal-Positive	Y		-0.194***		-0.129*
			(0.037)		(0.059)
Vocal-Negative	Y		0.366**		-0.074**
			(0.111)		(0.023)
Vocal-Arousal	Y		-0.037		-0.034
			(0.049)		(0.089)
Vocal-Valence	Y		-0.154		-0.102
			(0.160)		(0.125)
Verbal (Text)					
Verbal-Positive			0.030		-0.070
			(0.110)		(0.102)
Verbal-Negative	Y		0.045		0.000
			(0.281)		(0.131)
Verbal-Warmth	Y		-0.409***		-0.148***
			(0.093)		(0.027)
Verbal-Ability	Y		0.258***		-0.070
			(0.069)		(0.041)
Observations		96	96	192	192
Age Controls		Y	Y	Y	Y
Firm/Team Controls		Y	Y	Y	Y
Accelerator + Ind FE		Y	Y	Y	Y

Notes: OLS Regression. The analysis is obtained using the following model:

$$Performance = \beta_{vocal} \cdot X_{vocal} + \beta_{vision} \cdot X_{vision} + \beta_{text} \cdot X_{text} + \alpha_{FE} + \varepsilon.$$

The total employment and total funding raised are of the startup as of mid-2019. Both dependent variables are transformed using an inverse hyperbolic sine function. All pitch feature variables are standardized into a zero-mean variable with a standard deviation of 1. Standard errors clustered at the industry-level are displayed in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

Table 6. Gender Differences in the Pitch-Investment Relation

	(1)	(2)	(3)	(4)
	Dependent Var: *I(Invested)*			
	Man-only Teams		Woman-only Teams	
Pitch-Factor	0.019**		0.141***	
	(0.007)		(0.054)	
Visual (Facial)				
Visual-Positive		0.014***		0.052**
		(0.005)		(0.024)
Visual-Negative		-0.033***		-0.053
		(0.008)		(0.038)
Visual-Beauty		0.013*		0.061
		(0.007)		(0.047)
Vocal (Audio)				
Vocal-Positive		0.003		0.060
		(0.007)		(0.037)
Vocal-Negative		-0.040**		-0.604***
		(0.020)		(0.229)
Vocal-Arousal		0.015		0.095**
		(0.010)		(0.043)
Vocal-Valence		0.013**		0.107**
		(0.005)		(0.051)
Verbal (Text)				
Verbal-Positive		-0.005		0.050**
		(0.009)		(0.024)
Verbal-Negative		-0.020***		-0.107*
		(0.007)		(0.055)
Verbal-Warmth		-0.005		0.061*
		(0.008)		(0.031)
Verbal-Ability		0.004		0.017
		(0.011)		(0.027)
Observations	559	559	310	310
Video Characteristics	Y	Y	Y	Y
Firm/Team Controls	Y	Y	Y	Y
Accelerator FE	Y	Y	Y	Y

Notes: Logistic regressions, marginal effect. Separate analysis for man-only and women-only startup teams. The analysis is obtained using the following model:

$$I(Invested) = \beta_{vocal} \cdot X_{vocal} + \beta_{vision} \cdot X_{vision} + \beta_{text} \cdot X_{text} + \alpha_{FE} + \varepsilon.$$

$I(Invested)$ takes value of one if the startup team was chosen by the accelerator, and 0 otherwise. All pitch feature variables are standardized into a zero-mean variable with a standard deviation of 1. Standard errors clustered at the accelerator and year level are displayed in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

Table 7. Gender Differences in Gender-Mixed Teams

	(1)	(2)	(3)	(4)
	Dependent Var: *I(Invested)*			
	Man in Mixed-Gender Teams		Woman in Mixed-Gender Teams	
Pitch-Factor	0.044*		0.016	
	(0.024)		(0.037)	
Visual (Facial)				
Visual-Positive		0.035***		0.000
		(0.011)		(0.012)
Visual-Negative		-0.129*		0.005
		(0.074)		(0.008)
Visual-Beauty		0.007		-0.012
		(0.014)		(0.013)
Vocal (Audio)				
Vocal-Positive		-0.148		0.101
		(0.148)		(0.148)
Vocal-Negative		-0.355**		0.373**
		(0.170)		(0.168)
Vocal-Arousal		0.658***		-0.624***
		(0.215)		(0.209)
Vocal-Valence		0.368**		-0.286**
		(0.145)		(0.133)
Verbal (Text)				
Verbal-Positive		-0.492***		0.446***
		(0.156)		(0.138)
Verbal-Negative		-0.175		0.203
		(0.165)		(0.205)
Verbal-Warmth		0.001		0.018
		(0.104)		(0.096)
Verbal-Ability		0.035		-0.029
		(0.095)		(0.099)
Observations	270	270	270	270
Video Characteristics	Y	Y	Y	Y
Firm/Team Controls	Y	Y	Y	Y
Accelerator FE	Y	Y	Y	Y

Notes: Logistic regressions, marginal effect. Analysis for mixed-gender startup teams. The measures are calculated for female and male team members and incorporated into the model simultaneously:

$$I(Invested) = \beta_{vocal} \cdot X_{vocal} + \beta_{vision} \cdot X_{vision} + \beta_{text} \cdot X_{text} + \alpha_{FE} + \varepsilon.$$

$I(Invested)$ takes value of one if the startup team was chosen by the accelerator, and 0 otherwise. All pitch feature variables are standardized into a zero-mean variable with a standard deviation of 1. Standard errors clustered at the accelerator and year level are displayed in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

Table 8. The Informational Content of Human Interactions

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent Var: *I(Invested)*					
Measure	Unweight (Table 4)		First 5 Sec		TF-IDF	
Pitch-Factor	0.029***		0.025***		0.029***	
	(0.007)		(0.008)		(0.007)	
Visual (Facial)						
Visual-Positive		0.013**		0.023***		0.015***
		(0.006)		(0.006)		(0.005)
Visual-Negative		-0.029***		-0.014**		-0.027***
		(0.007)		(0.006)		(0.007)
Visual-Beauty		0.015**		0.011*		0.017***
		(0.006)		-0.006		(0.006)
Vocal (Audio)						
Vocal-Positive		0.011*		0.011		0.010*
		(0.006)		(0.007)		(0.005)
Vocal-Negative		-0.047***		-0.012*		-0.044**
		(0.017)		(0.007)		(0.019)
Vocal-Arousal		0.019**		0.018**		0.021**
		(0.008)		(0.007)		(0.008)
Vocal-Valence		0.020***		0.017**		0.021***
		(0.007)		(0.009)		(0.006)
Verbal (Text)						
Verbal-Positive		-0.011		-0.001		-0.011
		(0.009)		(0.004)		(0.007)
Verbal-Negative		-0.023***		0.000		-0.026***
		(0.007)		(0.008)		(0.010)
Verbal-Warmth		0.027***		-0.003		-0.012*
		(0.008)		(0.010)		(0.006)
Verbal-Ability		-0.043***		0.012*		0.003
		(0.007)		(0.006)		(0.009)
Observations	1,139	1,139	1,139	1,139	1,139	1,139
Video Characteristics	Y	Y	Y	Y	Y	Y
Firm/Team Controls	Y	Y	Y	Y	Y	Y
Accelerator FE	Y	Y	Y	Y	Y	Y

Notes: Logistic regressions, marginal effect. Measures are modified by accounting five seconds only, and by accounting for contents using TF-IDF. The analysis is obtained using the following model:

$$I(Invested) = \beta_{vocal} \cdot X_{vocal} + \beta_{vision} \cdot X_{vision} + \beta_{text} \cdot X_{text} + \alpha_{FE} + \varepsilon.$$

$I(Invested)$ takes value of one if the startup team was chosen by the accelerator, and 0 otherwise. All pitch feature variables are standardized into a zero-mean variable with a standard deviation of 1. Standard errors clustered at the accelerator and year level are displayed in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

Appendix

A. Collecting Video Data

We use an automatic searching script for two public video-sharing websites, YouTube and Vimeo. Integrated with query APIs, our web crawler returns a list of video indices according to a lot of pre-defined keywords, which include but not limit to the names of these accelerators, "startup accelerator application video","accelerator application videos" and so on. We first obtain the full list of potential videos returned by each keyword search (there is a limitation of returned videos by YouTube), and then filter the potential videos by a combination of different conditions on video info obtained along with the video itself. Filtering variables include but not limit to data format, duration, title, and annotation. It should be noted that because Y Combinator has too many application videos each year, we specifically specify the years in the searching keywords so as to obtain more complete results.

[Insert Table A.1 Here.]

We obtain additional videos from incubators' website. For example, MassChallenge provides more detailed information of alumni startups including company introduction and application video. Therefore, we also obtained these videos directly listed on MassChallenge's website as videos searching results and applied the filter stated above to expand our video dataset. However, notice that a proportion of videos provided by MassChallenge are application videos but advertisements, promotions or pitching videos, which have been excluded. We also employ startup names listed on incubators webpages to expand our video dataset. Specifically, we first obtain the full list of startups accelerated by the incubator each year if such a list is published on the incubator's website. Then our script automatically searches these startup names and checks the first three results returned by the search API. A match is defined as both the startup name and the accelerator name should appear in the video title or annotation.

It is worth noting that if one company has more than one videos in our sample, we only keep the video recorded first. There are 33 such firms in our analysis, which consist only

2.90% of our sample. These firms have multiple videos because of the following reasons. First, there are many entrepreneur teams applying for different accelerators. Second, there are some teams which applied for the same accelerator multiple times. For these firms, we only keep their videos and outcomes in the first application.

Figure A.1. Screen Shot of Search Results from YouTube

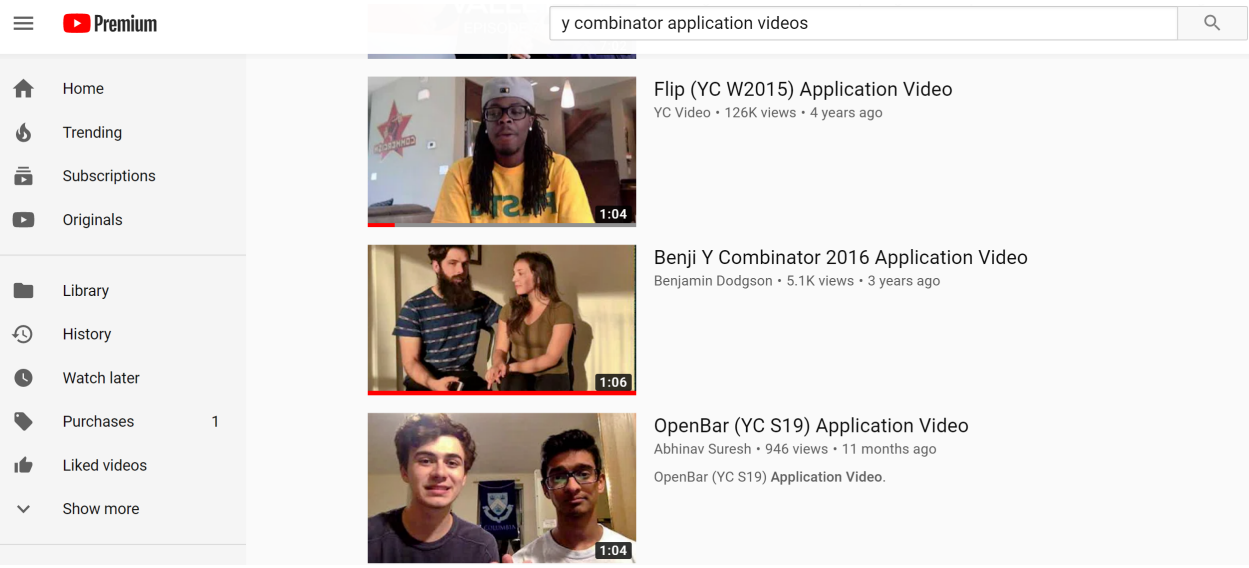


Table A.1. List of Searching Keywords for Collecting Videos

Keywords
YC Application Videos
Y Combinator Application Videos
MassChallenge Application Videos
500 Startups Application Videos
Techstars Application Videos
AngelPad Application Videos
Y Combinator Application Videos + *YEAR*
Techstars Application Videos + *YEAR*
500 Startups Application Videos + *YEAR*
AngelPad Application Videos + *YEAR*

Notes: This table shows the list of keywords we use for searching and collecting the pitch videos from Youtube and Vimeo. The *YEAR* takes values from 2005 to 2019.

B. Appendix Figures and Tables

Table A.2. List of Metrics from ML-based Video Processing and Collecting

Variable	Definition and Construction
A. Visual Metrics	
Visual-Positive	Probability of the facial emotion to be happiness by Face++ emotion recognition API
Visual-Negative	Sum of probabilities of the facial emotion to be sadness, anger, fear, and disgust by Face++ emotion recognition API
Visual-Beauty	Beauty scores for the faces in videos by Face++ beauty score API
B. Vocal Metrics	
Vocal-Positive	Probability of the vocal emotion to be happiness by the LSTM model in `speechemotionrecognition`
Vocal-Negative	Probability of the vocal emotion to be sadness by the LSTM model in `speechemotionrecognition`
Vocal-Arousal	Degree of the vocal arousal by the SVM model in `pyAudioAnalysis`
Vocal-Valence	Degree of the vocal valence by the SVM model in `pyAudioAnalysis`
C. Verbal Metrics	
Verbal-Positive	Whether a word is included in the positive category of the LM master dictionary
Verbal-Negative	Whether a word is included in the negative category of the LM master dictionary
Verbal-Ability	The direction (-1 or +1) of a word if it is included in the ability category of the NBS dictionary
Verbal-Warmth	The direction (-1 or +1) of a word if it is included in the warmth category of the NBS dictionary
D. Startup-level Variables	
I(Invested)	Whether the startup team is invested by the incubator
Total Financing Amt	The inverse hyperbolic sine transformation of the total amount of raised funding recorded in Crunchbase
Employment	The inverse hyperbolic sine transformation of the number of employees recorded in Pitchbook

Table A.3. Summary Statistics of Speaking Time Distribution in Pitch Videos

	Mean	STD	25%	50%	75%
Full Sample					
% Time of Speaking per Member	58.80%	33.72%	29.82%	57.20%	100.00%
# Times of Switching Speakers	2.96	3.50	0	2	5
Conditional on Multi-member Team					
% Time of Speaking per Member	43.05%	25.97%	22.86%	40.46%	62.39%
# Times of Switching Speakers	5.50	2.97	3	5	7
Conditional on Female-only Multi-member Team					
% Time of Speaking per Member	50.00%	19.48%	34.28%	50.00%	65.72%
# Times of Switching Speakers	5.91	3.50	4	6	6
Conditional on Male-only Multi-member Team					
% Time of Speaking per Member	43.18%	26.69%	22.71%	40.19%	62.62%
# Times of Switching Speakers	5.46	2.94	3	5	7
Conditional on Mixed Multi-member Team					
% Time of Speaking per Member	41.23%	24.86%	21.17%	39.98%	60.35%
% Time of Female Speaking	46.97%	26.34%	26.19%	48.36%	66.06%
# Times of Switching Speakers	5.53	2.96	4	5	6

Notes: This table provides summary statistics of the speaking time distribution in each pitch videos. The key variables of interests are the time of speaking per member and the time of switching speakers in each pitch. These characteristics are reported for full sample, multi-member team, women-only multi-member team, men-only multi-member team, and mixed-gender multi-member team. The speaking time variables are generated using the Speech-to-Text algorithm provided by Google.