Learning from machines: can CEO vocal cues help to predict future firm performance?

Zihao LIU

August 12, 2024

Abstract

I apply tailored deep learning models on CEO voiceprints of earnings conference calls to predict the firm's future performance as measured by analyst recommendation consensus changes, unexpected earnings, and cumulative abnormal returns. The out-of-sample evaluations of the models consistently outperform established benchmark models using only textual information and firm characteristics by approximately 13% on average, achieving a prediction accuracy ranging from 54% to 65%. In other words, how firm information is communicated in addition to the content can affect the market perception of a firm. This study adds new evidence to audio recordings of conference calls containing valuable information about a firm's fundamentals, incremental to qualitative "soft" information conveyed by textual content, and quantitative earnings information. To achieve some level of explainability of the learned voiceprints' nuances, I employ a tailored vocal emotion classifier, contributing to improving and refining vocal sentiment analysis in the existing literature (*JEL* D83, G14, G41).

Keywords: Deep learning, Transfer learning, CEO, Earnings conference calls, Measuring sentiment, Multimodal models

1 Introduction

Economic decisions without any form of human interaction are uncommon. For example, when a financial analyst evaluates a firm's future perspectives, the choice to recommend buying or selling a stake in the firm will consider the fundamentals such as firm characteristics, past performance, and corporate governance. In addition to classical verbal information (textual content) such as earnings reports, analysts incorporate nonverbal communication (paralinguistic features) in the conference call - the interaction that predates the analyst recommendations or analyst reports to make their economic decisions (Barcellos and Kadous, 2022; Choudhury et al., 2019; Mayew and Venkatachalam, 2012). Therefore, it is interesting to investigate how the non-verbal behaviors of the CEO explain analysts' or investors' behaviors, whether analysts or investors respond to a CEO's vocal features, whether the response is rational or irrational, and which emotional characteristics positively or negatively affect investors' decisions. Similarly, many other decisions are also made based on human interactions - for example, company interviews for hiring and roadshows of initial public offerings.

It can be hypothesized that financial decisions can be improved by considering information about interactions (Hirshleifer, 2020; Hu and Ma, 2021). Economic agents mostly make decisions with imperfect information. Interactions among these economic agents provide additional information on characteristics such as leadership ability or confidence in the firm, which may help agents form informed beliefs and make better decisions. On the other hand, human interactions may originate or augment biases. These biases could be preference-based, i.e., agents receive non-pecuniary utility from interacting with or contributing to people with certain features revealed in interactions (e.g., "I enjoy listening to energetic people"). Alternatively, these biases could be due to erroneous beliefs - agents incorrectly form beliefs based on characteristics obtained from human interactions (e.g., "This investment advisor sounds/looks passionate, so the advice must be good").

Understanding how interactive communication shapes such choices is challenging, mainly due to the limitations of traditional econometric methodologies and structured data formats. Given the importance of human interactions in many essential economic decisions, I harness the capabilities of artificial intelligence (AI) technology and alternative data, such as audio recordings, to research human interactions. This paper contributes to vocal interactions, a crucial aspect of human engagement. First, I focus on empirically measuring vocal features using spectrograms (voiceprints) in a corporate setting. Second, I test whether vocal communication features have economic consequences using deep learning

(DL) models. Third, I employ multimodal models to integrate numerical representation (embeddings) of vocal features with textual sentiment features and firm characteristics to predict future firm performance more accurately.

This paper answers these questions and makes progress on three fronts. First, I propose a tailored method to use raw audio as data input. This involves employing a DL model to transform audio recordings from vocal interactions into numerical representations of paralinguistic features of vocal cues, e.g., emotion and tone, from human interactions. This method offers significant flexibility in developing and customizing algorithms and measurements, allowing us to answer specific research questions. For instance, these paralinguistic features can be utilized for refining emotions classification like the audio sentiment analysis employed in Gorodnichenko et al. (2023) and Hu and Ma (2021). Additionally, I can use the extracted paralinguistic features to classify whether a CEO is depressed using DL models developed in Homsiang et al. (2022).

Second, I apply the tailored methodology to CEO speeches in quarterly earnings conference calls of publicly listed companies of the S&P 500 index. These conference calls are mostly audio-only, which allows us to isolate the audio channel from the other channels, such as facial expressions and gestures (Momtaz, 2021; Curti and Kazinnik, 2021; Hu and Ma, 2021). An earnings conference call typically has two parts that involve the CEO (or other executives) - the executives' prepared speech and the improvised answers to questions from analysts in the question and answer (Q&A) section. A higher positive perception of speeches could be, for example, reflected in sentiment features such as confidence, passion, and warmth, which are associated with a higher probability of buying recommendations from financial analysts. I differ from Gorodnichenko et al. (2023) and Hu and Ma (2021) by using pre-trained DL models to extract vocal features directly associated with analyst recommendation consensus changes immediately after the earnings calls that are objectively observed. My models do not require subjectively defined features associated with speech emotions, focusing instead on learning directly from DL models without discarding valuable information[1]. In a later stage, I try partially opening the "black box" using emotion classifiers to understand the economic interpretations of extracted vocal features.

Third, I further investigate whether my vocal sentiment features add value to the textual sentiment features (Loughran and McDonald, 2011, 2020) and firm characteristics (Fama and French, 2015; Ehsani and Linnainmaa, 2022) using multimodal models

[1]The speech emotion recognition algorithms are trained based on a few actors' performances (Burkhardt et al., 2005; Livingstone and Russo, 2018) and the speech emotions are not yet well defined in the literature (Lacerda, 2012; Schuller, 2018).

(Wankhade et al., 2022). By incorporating vocal embeddings (numerical representation of vocal features), the multimodal models outperform the benchmark models, which only use textual sentiment features and firm fundamentals. Moreover, I further validate my models using K-fold cross-validation and show the average prediction results with confidence intervals instead of focusing on the single best-performed model. This is the first paper to apply K-fold cross-validation on such multimodal models in finance.

I analyze audio data using vocal information as spectrograms[2] (voiceprints) and textual content of verbal information, including transitory information, such as confidence and sentiments, and persistent information, such as the CEO's charisma and persuasiveness. I map the unstructured raw audio data into numerical representations to obtain a data format suitable for deep learning algorithms. Additionally, I employ transfer learning, enabling the use of pre-trained image classification models for extracting audio features, as demonstrated by Beckmann et al. (2019).

To be more specific on the input data, I represent the audio information as spectrograms of the CEO speech's second sentence and extract the speech's textual information using earnings conference call transcripts and speech-to-text algorithms. The second sentence is chosen because the speech usually starts from the second one. In contrast, the first sentence is mainly sentences like "Welcome, everyone.". Furthermore, in terms of the length of audio recordings, I use a 5-second length. The deep learning models exhibit strong performance across various prediction tasks, even with a brief audio duration. For instance, 2-second speech audio proves sufficient for accurately detecting emotions, Dysarthria, or mask-wearing, as demonstrated by Shor et al. (2022). The results are robust for the models using first sentences, a random selection of sentences, or 4-second segments of the audio recordings.

I use well-established DL methods, mainly convolutional neural networks (CNNs)[3], to map the speech spectrograms onto predictions of economic performances. This approach allows for replicability, transparency, and reasonable research computation burdens. In this paper, I apply two different strategies by using CNNs. First, I train DL models from scratch and apply simple model architectures with one or two layers of CNNs. Secondly, I

[2]A spectrogram is a visual representation of the spectrum of frequencies in a signal as they vary with time. It is a way to analyze and display the frequency content of a time-varying signal, such as audio or other time-domain data. Spectrograms, focusing on audio information extraction, are a widely utilized format in deep learning (DL) models for,e.g., audio signal processing, speech analysis, and music analysis.

[3]A convolutional neural network (CNN) is a regularized form of a feed-forward neural network, autonomously acquiring feature engineering skills through the optimization of filters or kernels. I show CNN structure in the Online Appendix.

use transfer learning and tailor the pre-train model architectures - VGG16 (Simonyan and Zisserman, 2014) and SpeechVGG (Beckmann et al., 2019) - to adopt the original models into the context of finance. I use the multi-dimensional outputs from the second last layer of my tailored DL models to predict and construct measures on vocal features. The extracted audio information can include non-verbal cues, e.g., laughter, hesitations, and consent. But, again, I am not using these non-verbal cues labels directly in my models. I use the raw audio information as this contains the most granular information. To combine measures of vocal features with textual characteristics and financial fundamentals, I apply the multimodal deep regression model (MDRM) to incorporate different dimensions of communication features - verbal, vocal, and quantitative.(Qin and Yang, 2019; Hu and Ma, 2021; Wankhade et al., 2022; Gorodnichenko et al., 2023). By combining information from multiple modalities, my multi-modal models incorporating audio information can make more accurate predictions than benchmark models with textual information (Loughran and McDonald, 2011) and firm characteristics (Fama and French, 2015; Ehsani and Linnainmaa, 2022).

I find that features in human interactions are closely related to economic consequences. My results show that the spectrograms of a CEO speech have additional predictive power on future firm performance measured by analyst recommendation consensus changes, unexpected earnings, cumulative abnormal returns (CAR(0,2)s, and CAR(2,127)s) using my transfer-learned model based on SpeechVGG (Beckmann et al., 2019). The multi-dimensional audio factor model has an average prediction accuracy of 62.2%. This improves the prediction accuracy by 12.5% on average from the benchmark model using textual information and financial factors (e.g., firm size, book-to-market ratio, historical return volatility, and momentum). Moreover, the prediction accuracy on long-term (half-year) firm performance measured by CAR(2, 127) is the highest (65.1%) on average.

An economist reader might ask herself the question: "Why do speech features matter for future firm performance?" The previous results suggest that investors respond to speech features. The next question is whether this response is rational (i.e., these features reveal additional information about firm fundamentals) or irrational (i.e., they reflect emotional bias). Therefore, I partition potential explanations into two broad categories: explanations where investors have interaction-induced biases and explanations where they do not. Without interaction-induced biases, investors maximize a purely pecuniary objective function, and interactions provide information to help calibrate their beliefs about the "quality" of the firm. For example, being positive and passionate might be desirable and success-

enhancing personality traits for corporate governance. Alternatively, communication and interpersonal skills may be productive in leading a firm.

In contrast, explanations involving interaction-induced biases argue that investors' favoritism for certain interactive features may arise from taste-based reasons and inaccurate beliefs. Positivity demonstrated in pitches may be particularly salient in leading to these biases through affecting investors' emotional states (DellaVigna, 2009). Emotions and moods are contagious, so more positive pitches help investors achieve a favored positive emotional state as a utility. Moreover, emotions and moods influence both beliefs about prospects and assessments of risk. These effects are shown in many economic settings, notably when factual information lacks (DellaVigna and Pollet, 2009).

I investigate explanations without interaction-induced biases by exploiting the firm's long-term performance with a positive short-term performance measured by CAR(0,2)s (Ewens and Townsend, 2020). Under unbiased belief calibration, the companies with positive stock reactions and higher levels of speech positivity would likely perform better than those with inferior speech features. I track firm performance using profitability (earnings per share) and half-year cumulative abnormal returns (CAR(2,127)s). I find positive speech features are linked to the firm's better long-term performance and show evidence of no interaction-induced biases. Thus, the fact that the prediction accuracy on long-term (half-year) firm performance measured by CAR(2, 127) is the highest shows a rejection of interaction-induced biases[4].

Finally, to achieve some level of explainability of the learned embeddings, I developed a vocal emotion classifier using the same pre-trained architecture. Subsequently, applying the emotion classifier to the CEO vocal embeddings revealed that a positive change in analyst evaluations is reflected in vocal patterns that are more happy and, surprisingly, more fearful. We conclude that current deep vocal embeddings support the prediction of analyst recommendations and that some explainability can be achieved by a learned

[4]The investigations on economic mechanisms are ongoing. In the final part of the paper, I will further explore the source of this bias using an investment experiment. I follow the structure of Bohren et al. (2019) and aim to distinguish and quantify inaccurate beliefs versus taste-based channels. The subjects, finance students with basic finance and investment training, are shown 10 synthesized CEO speeches based on the trained AI models and asked to make investment choices about those firms to maximize their payoff, which in turn depends on the real-world performance of the firms. Following the previous analyses, I expect that subjects are more likely to invest in firms with more positive speech features.

Notably, the experiment elicits the beliefs of the subject investors. Consistent with the inaccurate beliefs channel, I expect that investors mistakenly think that firms with more positive speech features are more likely to succeed, even though those companies' realized performance is lower, as discussed above—hence the inaccurate beliefs. On the other hand, I would find that speech features remain influential as a standalone determinant after controlling for the investment decision's elicited belief; then, the results are consistent with the taste-based channel.

mapping of the embeddings onto emotions.

This paper proceeds as follows. Section 2 discusses the background and the contribution to the related literature. Section 3 documents my datasets and empirical measurements (firm performance and control variables). Section 4 discusses the audio features and deep learning models. Furthermore, Section 5 shows my model selection processes and prediction results based on CEO speeches. Moreover, Section 6 tries to explain the embeddings of CEO speeches partially. Finally, I conclude in Section 7 with economic interpretations and develop plans for further research.

2 Backgrounds and Related Literature

This paper contributes to the growing literature, including but not limited to (i) social economics and finance and behavioral finance, (ii) artificial intelligence (AI) and big data in economic research, and (iii) manager characteristics and corporate governance. Economic and finance theories are built on what agents know and how agents perceive the acquired information. Meanwhile, behavioral finance is an active area that accommodates investors' psychology, cognition, and irrational decision-making. As an extension to standard information economics, where people interact only via signals such as trading orders and observation of information such as market price, social economic and finance theories incorporate the importance of human interactions. Hirshleifer (2020) emphasizes the social interactions that shape economic thinking and behavior. The economic behaviors of agents involve both in-personal interactions and communication. As a result, this research strand will require empirical measures of unstructured data such as audio data, text contents, and graphical information. Luckily, the emerging modern technologies using AI make studies on human interactions possible (Liebregts et al., 2020).

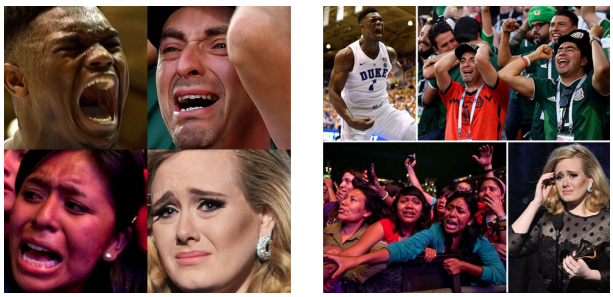
As a relatively mature method, since Tetlock (2007), the literature in Finance and Accounting studying different types and analysis methods of textual data has flourished (Loughran and McDonald, 2020; Garcia et al., 2023). Recent textual-analysis studies have also shifted from traditional media sources such as newspaper articles (Engelberg and Parsons, 2011; Tetlock, 2007) to more interactive information channels (Campbell and Shang, 2022; Long and Zhong, 2023). Social media is an interactive platform that facilitates information exchange by enabling dynamic and often repeated user interactions that affect market activities. For example, in an experimental setting, Elliott et al. (2018) shows that investors trust the CEO more and are more willing to invest in the firm when the CEO

communicates directly through a personal Twitter account. This communication helps investors develop a more robust social bond with and trust in the CEO. Observing how the information is communicated is essential if it is less important than the information content itself.

While the text contents of messages, company announcements, financial reports, and press releases (among others) are investigated quite well Loughran and McDonald (2011, 2020), there still needs to be more empirical economic and financial research on in-person interactions critical in individual decision-making in addition to verbal interactions. People observe extra information such as appearance, facial expressions, and gestures and acquire acoustic signals by meeting and talking to each other (Duarte et al., 2012; Blankespoor et al., 2017; Graham et al., 2017). However, social psychology research suggests that vocal indicators of various emotions or sentiments are accurately detected and often as good or better than facial cues and expressions (Kappas et al., 1991). It is also interesting to check the following example in Figure 1 on why facial expressions are complex to interpret, even for humans.

Why Facial Expressions and Gestures are Hard to Interpret?

Figure 1: This figure presents the happy facial expressions of four people under four different circumstances. The picture group on the left-hand side focuses only on faces, while the group on the right zooms out by incorporating gestures. Anti-clockwise, from top left in both groups: basketball player Zion Williamson celebrates a dunk; Justin Bieber fans cry at a concert in Mexico City; singer Adele won Album of the Year at the Grammys in 2012; Mexico fans celebrate a win in a World Cup group match(Heaven, 2020).



Even though Figure 1 shows the happy moments, facial expressions look sad, and gestures sometimes do not help capture the facts. Therefore, it is economically meaningful to investigate non-verbal information by focusing on vocal cues or combining them with other non-verbal communication. This strand of literature extends the textual analyses by exploring the non-text data such as the vocal cues of managers in earnings conference calls (Mayew and Venkatachalam, 2012), videos (graphs and audio signals over time) of

startup pitch presentations (Hu and Ma, 2021) and speech emotions of federal reserve chairmen or chairwomen in Federal Open Market Committee (FOMC) press conferences (Gorodnichenko et al., 2023). This research strand is still in an early stage and is often criticized due to the "black boxes" of commercial software used in the research and the inconclusive measurement of affective states. For example, Lacerda (2012) argues that the commercial software [5] that measures the affective states of managers in Mayew and Venkatachalam (2012) is eventually useless and generates pure noise.

That said, Mayew and Venkatachalam (2012) still established the suggestive importance of nonverbal vocal communication of executives in conference earnings calls. Hu and Ma (2021) also relies on speech emotion recognition algorithms to measure a startup pitch's positivity by aggregating all text, audio, and graphical information into one factor, making the economic interpretation hard to achieve. This work addresses the underlying causal factors of nonverbal vocal communication and economic mechanisms. Similar to Gorodnichenko et al. (2023), in this paper, I focus on vocal cues instead of facial expressions and gestures. However, the methodologies of this paper refine the basic deep learning model applied in Gorodnichenko et al. (2023) and employ transfer learning models. Transfer learning in deep learning is a methodology that employs pre-trained neural network models as a foundation for training new models on tasks that are distinct yet interconnected. I will discuss more details regarding transfer learning in Section 4 about methodologies.

Moreover, I focus on out-of-sample predictions in a corporate setting instead of impulse response analysis on macroeconomic indicators compared to Gorodnichenko et al. (2023). My paper's multimodal models are also validated using K-fold cross-validation. Each model's average prediction accuracy and confidence intervals are reported and discussed instead of focusing on a single best-performed model. Cohen et al. (2020), and Mayew (2008) find that there are firms "cast" their conference calls by calling on bullish analysts disproportionately in the question-and-answer (Q&A) sections of earnings conference calls [6]. These firms tend to underperform their peers in the future.

This paper introduces new empirical methods and evidence to extend the existing literature's frontiers. Firstly, this paper offers new empirical measurements of non-verbal cues, such as vocal sentiment, by opening the "black boxes". Second, I try to disentangle the casual human interactive factors in speeches that affect economic and financial decision-making. Third, I test why vocal cues and interactions matter for economic and financial

[5] https://www.nemesysco.com/lva-technology/

[6] The Q&A data is available in the newly collected dataset and needs to be added to this version. However, the preliminary results in Liu et al. (2023) show that the prediction accuracies based on Q&A sessions are not very different from the results based on CEO-prepared speeches.

decisions, whether these human interactions help agents improve decision-making by overcoming information asymmetry, or whether these interactions lead to different behavioral biases.

The newly developed methodology in this work is required to achieve the above-mentioned exploration. The new method lies in an emerging research line that deals with unstructured data and AI techniques for economic and financial research. As I briefly discussed textual analysis studies, Gentzkow et al. (2019) and Loughran and McDonald (2020) summarize textual data analysis in economics and finance, while Bochkay et al. (2023) summarizes the accounting literature. To my knowledge, this work's approach provides the first thorough exploration of a manager's speech data by using raw audio data and spectrograms (voiceprints) of a manager's speeches during an earnings conference call and deep learning models based on CNNs. My DL-based method has several advantages. First, this method does not rely on a consistent measure of affective states, which still needs to be developed, but tries to use objectively observed analyst recommendation consensus changes immediately after the earnings conference calls.

Moreover, the non-verbal cues are measured by directly analyzing the spectrograms of speeches. This choice compromises the accuracy and computation power between analyzing raw audio data directly (Oord et al., 2016) and analyzing audio features generated from spectrograms such as Mel-frequency cepstrum coefficients (MFCCs) and pitches (DellaVigna and Pollet, 2009; Qin and Yang, 2019). The transparency of my measures improves the reliability and reproducibility of empirical results. Finally, the analytical framework and procedures are scalable to incorporate more structured and unstructured measures in different research settings by using MDRM.

The paper's research settings relate to the literature on CEO characteristics and corporate governance, a central topic in corporate finance. Observed CEO personality traits (Barcellos and Kadous, 2022; Kaplan et al., 2012; Malmendier and Tate, 2005) and self-reported management style (Mullins and Schoar, 2016) affect firm performance as well as investors' decisions. Investors may rationally value manager abilities, such as leadership (Bandiera et al., 2020) and work experience (Giannetti et al., 2015; Custódio and Metzger, 2014). However, investors can also be biased by discriminating factors such as gender and race (Ewens, 2022; Francis et al., 2021). This paper aims to contribute to the existing literature by providing a detailed analysis and potentially large-sample study on the influence of human interactions in corporate settings.

3 Data

The paper uses quarterly earnings conference calls of companies included in the S&P 500 index as my laboratory. This paper uses a published dataset provided by Qin and Yang (2019). They separated the CEO speeches from original earnings conference calls of S&P 500 companies during the sample period between January 1 and December 31, 2017. The original audio files are available on the Thomson Reuters StreetEvents database. After merging with database CRSP, Compustat, and I/B/E/S for return, accounting, and analyst variables, 543 firm-quarter observations remain.

3.1 Audio data and earnings conference call

As a vital firm communication channel, earnings conference calls are investigated by extensive literature studying the relationship between firms and analysts and studies of the information content of earnings announcements and earnings conference calls (Baik et al., 2023; Cen et al., 2021; Suslava, 2021; Cohen et al., 2020; Mayew and Venkatachalam, 2012). While earnings conference call transcripts have been extensively studied in prior research, I focus on audio information in addition to textual analysis in this work. The textual transcripts of earnings calls are well-labeled, including the speaker identifications (executives and analysts) and speech content. On the other hand, the audio data downloaded directly from Thomson Reuters StreetEvents does not provide any segmentation or labeling for speakers within an earnings conference call. More earnings call audio and transcripts are available on the Thomson Reuters StreetEvents database.

To balance the precision and plausibility concerning extracting audio features, I analyze the earnings conference calls at the sentence level, i.e., a sequence of sentences represents a conference call with corresponding audio clips. The audio data is processed using the Iterative Forced Alignment (IFA) algorithm to align each transcript with the audio clip containing the corresponding spoken text in each sentence Qin and Yang (2019). Furthermore, I select only the speeches made by a firm's CEO and exclude the observations that the CFO gave speeches in the earnings conference calls. I use Python packages SciPy and Praat2 to extract vocal features, such as pitch, intensity, MFCCs, and spectrograms, from raw audio data.

I built my pilot dataset by acquiring all available S&P 500 companies' quarterly earnings conference calls in 2017 published by Qin and Yang (2019). They choose S&P 500 constituent firms as the target for firm performance prediction due to a combination of com-

pany importance and tractability. Firms in the S&P 500 index comprise approximately three-quarters of the total U.S. market capitalization by the end of 2017. In total, the dataset contains 2,243 quarterly earnings conference calls in 2017. The conference calls in which text-audio alignment is not done correctly are discarded using the above-mentioned data processing method and this results in a final dataset consists of 576 conference calls. I dropped 33 observations due to missing information on the CARs or Speeches not performed by a CEO. I eventually had 543 firm-quarter observations with a total number of 88,829 sentences. Much raw data is discarded because the audio-text alignment is noisy and prone to errors. The processed earnings conference calls (text and audio) are available online.

I expanded the datasets beyond the pilot data in the latest version of the empirical analysis. I collected 764 firm-year observations for 2019 and 2020 of S&P 500 companies and manually separated the CEO speeches and CEQ answers in the Q&A sessions from the raw earnings conference call recordings[7]. By using the manually collected datasets, Liu et al. (2023) developed a state-of-art deep learning model based on the TRILLsson model (Shor et al., 2022) to extract the paralinguistic features from CEO speeches for predicting analyst recommendation consensus changes.The preliminary results in Liu et al. (2023) further confirms the empirical findings in this paper by using a more recent deep learning model architecture to extract audio features. In the subsequent section (6), I will delve deeper into the discussion of these state-of-the-art deep learning models, which were trained on the most current datasets.

3.2 Firm information

Firm information is collected from widely used databases in economic and finance research. CRSP, Compustat, and I/B/E/S databases are used for stock return, accounting, and analyst data. I search for a company's financial information using the company's ticker with available audio data. By following Mayew and Venkatachalam (2012), I mainly measure the firm performance using changes in analyst recommendations, unexpected earnings (UEs), and cumulative abnormal returns CAR(0,2) for short-term firm performance. I use half-year (based on the business calendar) cumulative abnormal returns CAR(2,127) after the date of an earnings conference call for long-term firm performance. In Section 5, I will

[7]For now, I manually collected 105 firm-year audio and text files covering the Dow & Jones 30 index and the NASDAQ100 index in 2017 for out-of-sample evaluations. I can eventually acquire 38,393 firm-quarter observations in the U.S. and 11,356 in Europe from 2017 to 2021. The data universe keeps updating over time. However, the eldest datasets are deleted from the StreetEvents database due to the ample storage space.

also control for firm characteristics such as stock return momentum, book-to-market ratio, and firm size[8].

Sample Description of Firm Performance and Characteristics

Table 1: This table presents the mean, standard deviation (SD), minimum (Min), 25th quantile, median, 75th quantile, maximum (Max), and the number of observations of firm performance measures and firm characteristics for 543 earnings conference calls of S&P 500 companies in 2017. The ΔRECs measures changes in the difference between the percentage of buying recommendations and the percentage of selling recommendations after the earnings conference call for each firm. The unexpected earnings (UE) score measures the standard deviations in the actual(reported) earnings that differ from the I/B/E/S surprise mean estimates for each firm after the earnings conference call. The EPS documents the quarterly earnings per share (EPS). The CAR(0,2) and CAR(2,127) variables stand for 3-day cumulative abnormal returns (CARs) and half-year CARs, respectively, estimated using the capital asset pricing model (CAPM) after the date of the earnings conference calls. Return volatility is the past half-year volatility of firm stock returns before the earnings conference call. The book-to-market ratio equals the book equity divided by the firm's market value. Market cap. ($ billion) is each firm's total market capitalization in billion USD dollars at the end of the earnings quarter. Ln(market cap.) is the natural logarithm of the total market capitalization of the firm at the end of the earnings quarter.

	Mean	SD	Min	25th	Median	75th	Max	N
ΔREC	0.10	4.66	-20.51	-1.24	0.00	1.10	16.66	543
UE score	1.92	4.02	-29.22	0.26	1.37	2.85	34.02	543
Average UE	1.04	0.84	-3.91	0.55	0.86	1.36	5.04	543
CAR(0,2)	-0.00	0.06	-0.29	-0.03	-0.00	0.03	0.25	543
CAR(2,127)	0.02	0.27	-0.88	-0.13	-0.01	0.17	1.15	543
EPS	1.09	0.88	-4.00	0.57	0.89	1.46	5.35	543
Return Volatility	0.01	0.01	0.01	0.01	0.01	0.02	0.04	543
Book-to-market ratio	0.30	0.25	-0.25	0.12	0.25	0.41	1.45	543
Market cap. ($ billions)	47.99	69.15	1.88	11.76	23.20	48.79	531.31	543
Ln(Market cap.)	10.19	1.01	7.54	9.37	10.05	10.80	13.18	543

Table 1 displays the distribution of firm performances within our sample. Most firm performance variables exhibit a slight skew towards positive outcomes, as indicated by their mean values, except for CAR(0,2), which hovers around an average of approximately 0. It is important to note that the distribution of firm performance variables in our sample is fairly balanced. Our key variable of interest, changes in analyst recommendation consensus, defined as the percentage difference between buying recommendations and selling

[8]The CEO characteristics data will be added in the later version of the paper from the BoardEx database to explain economic mechanisms further.

recommendations by analysts following the same firm, is notable for its zero median and a mean of approximately 0.10%, which is close to zero when contrasted with its relatively wide standard deviation of 4.66%. Therefore, the distribution of firm performance is not a significant concern. Moreover, I will also train my benchmark models based on binary classification variables based on the zero threshold of a variable. For example, I will classify whether a CEO's speech is perceived as good or bad based on whether the changes in analyst recommendation consensus are above zero.

4 Methodology

The audio data document human behaviors of what people say and how they say it. My method of processing raw audio data includes three steps, as shown in Figure 2. First, I decompose the speech information into two parts - speech spectrograms and text contents. Secondly, I started the research by training DL models to directly predict analyst recommendation census changes by analyzing corresponding speech spectrograms. I incorporate the textual sentiment of earnings conference calls and firm characteristics in a multimodal model. We train linear (Logistic) and nonlinear (random forest (RF) and support vector machine (SVM)) models to predict future firm performance. For the Logistic model, I take the default settings, and for the RF and SVM models, I perform automatic tuning (using the training folds only). Finally, I adopt AI algorithms to create interpretable features from the raw audio data for each part by exploring the model weights and visualization in DL (Zhang et al., 2021; Yosinski et al., 2015; **?**)[9]. Below, I discuss the information transmission process of earnings conference calls, analysis of audio data relevant to economic research, and then describe the deep learning methodology in detail.

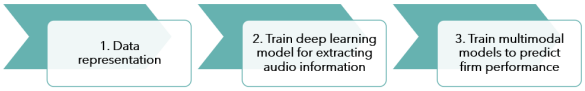
4.1 Information transmission mechanisms

In Figure 3, I present an illustrative depiction of the information dissemination process for the 2nd quarter of 2023 regarding Booking Holdings Inc.'s financial performance. In the context of training a deep learning model designed to extract non-verbal cues (paralinguistic features) from the CEO, I focus on two pivotal events, which are prominently highlighted in red boxes: the live stream of the earnings conference call and the subsequent

[9]The interactions with analysts in the Q&A sections are not yet in the dataset of this draft version.

Three Steps for Model Training

Figure 2: This figure presents three main steps of the model training process for predicting future firm performance based on audio, textual, and financial information. The first step - data representation - is to prepare the spectrograms of CEO speeches as model inputs. In the second step, I train multiple deep-learning models, including a transfer-learning model based on the pre-trained SpeechVGG model (Beckmann et al., 2019) to extract audio information from spectrograms. In the final step, I incorporate the trained model embeddings from the second step to train multimodal models to predict future firm performance, measured by analyst recommendation consensus changes, unexpected earnings, three-day cumulative abnormal returns, and half-year cumulative abnormal returns.



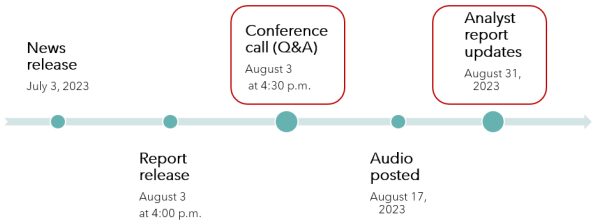
analyst recommendation updates provided in analyst reports.

It is worth noting that the live stream of earnings conference calls remains an inclusive platform accessible to all stakeholders. Nevertheless, the primary participants who engage directly with CEOs during these earnings calls are financial analysts, and their interactions are well documented in the earnings call transcripts. To this end, the initial focus of the deep learning model training involves extracting CEO non-verbal features from the shifts in analyst consensus recommendations. This choice is rooted in the assumption that these financial analysts serve as representatives of other stakeholders.

Indeed, the changes in analyst recommendations reflect the reactions and responses from CEO communications and interactions during these critical earnings conference calls and other written information such as news releases and financial reports. Therefore, by delving into this specific dataset, we aim to capture valuable insights into the nuanced paralinguistic elements of CEO communication that can significantly impact stakeholder perceptions and decision-making by controlling for textual content and other firm characteristics. Subsequently, I test market perceptions of the non-verbal cues using short-term and long-term cumulative abnormal returns. This evaluation aims to determine whether the CEO's paralinguistic features, derived from pre-trained deep learning models based on changes in analyst recommendations, add significant value alongside textual and financial information.

An example of information release process

Figure 3: This figure provides a comprehensive timeline of key information releases during the 2nd quarter of 2023 (2023Q2) of Booking Holdings Inc.'s financial performance. On July 3, 2023, the company issued a significant news release detailing the 2023Q2 financial reports and the earnings conference call. Subsequently, the actual financial reports were made publicly available on August 3, 2023, at precisely 4 p.m., a mere thirty minutes ahead of the commencement of the live stream for the earnings conference call, which was scheduled for 4:30 p.m.. This is 30 minutes before the closing time of the financial markets at 5:00 p.m.. It's noteworthy that the audio recording of this conference call became accessible two weeks after the live-stream event. Finally, the analyst reports, summarizing their assessments and opinions, were disseminated at the close of the corresponding month, specifically on August 31, 2023.



4.2 Audio features

One of the most fundamental questions is how to represent the audio data when designing a tailored DL model to process the CEO acoustic features. Various representations have been frequently applied in the literature, including hand-crafted audio features such as pitch and tone, Mel-frequency cepstral-based features such as the Mel-frequency cepstral coefficients (MFCCs) and spectrogram-based images, and raw audio data. While there is yet to be a consensus on the best input representation for a DL model (Purwins et al., 2019), I motivate my choice of using spectrogram-based features for my data representation.

The drawback of the hand-crafted features is that these designed features may not be optimal for the task objectives, and much information might be lost. In recent decades, the Mel-frequency cepstral coefficients (MFCCs)[10] have been the dominant feature representation for audio analysis tasks. One of the advantages of MFCCs is that they are a very sparse representation of the original audio. However, the non-linear transformation for generating MFCCs removes information and harms spatial relations, which is not optimal for DL models.

[10] "The MFCCs are magnitude spectra projected to a reduced set of frequency bands, converted to logarithmic magnitudes, and approximately whitened and compressed with a discrete cosine transform (DCT)"(Purwins et al., 2019)

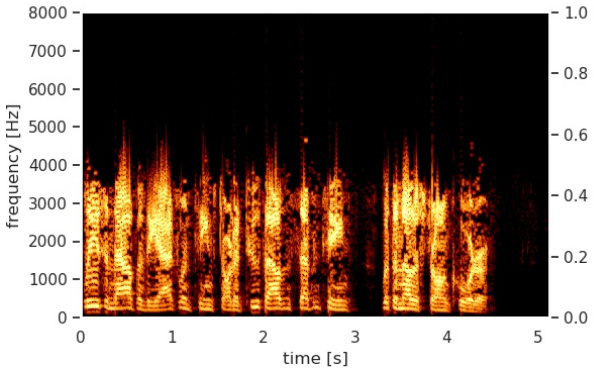
There has recently been a shift towards training models directly on the raw data. For example, DeepMind designed a convolutional architecture called WaveNet to generate audio (Mohamed et al., 2022; Oord et al., 2016). These WaveNet are trained on the raw audio, and not only can they be used for speech creation, speech recognition, and other classification tasks. It would be more optimal to train a DL model on more information than the MFCCs, but DL models like WaveNet can be computationally expensive to train and adopt.

Spectrograms (voiceprints) are relevant in this research because they retain much more information than MFCCs about the original audio signal and are computationally cheaper to train than DL models like WaveNet. A spectrogram is a temporal sequence of spectra. It is a graphical representation of audio with frequency on the vertical axis and time on the horizontal axis. The third dimension of color can be added to represent the acoustic intensity. However, the inputs into the deep learning models are ultimately the matrices of numerical values that generate spectrograms. Within these matrices, the columns represent the temporal dimension, capturing the progression of time, while the rows imply distinct frequency bands within the audible spectrum for humansconstrained between 10 Hz to 8,000 kHz for the inputs. Each numerical value encapsulates the intensity of the voice at a specific intersection of time and frequency. To ensure uniformity and comparability, all values in the matrices undergo z-score standardization. This process enhances the model's ability to extract meaningful patterns and relationships across observations, contributing to the efficacy of voice recognition and analysis. The model inputs do not include figure labels or axes. Figure 4 shows a used example of a speech spectrogram of Michael R. McMullen, the CEO of Agilent Technologies Inc., during the earnings conference call on February 14, 2017.

In this spectrogram, we can see many voice contours on different speech frequencies. These contours can extract text contents from the spectrograms in speech recognition tasks. The vertical dark lines are the unvoiced parts of speech (exact speech pitch can be detected) and brief pauses between words or breath sections (exact speech pitch is unavailable) in a sentence throughout the spectrogram. Therefore, we can see that the spectrogram contains much information about the nature of different speeches. The other advantage of using spectrograms as inputs is that we can use the AI algorithms of image classification, which has many breakthroughs and is a well-developed area in the AI literature. There have been many developments related to computer vision through advances in deep learning. Large datasets such as ImageNet for training deep learning models (Deng et al., 2009),

A Used Example of CEO Speech

Figure 4: This figure presents the spectrogram (voiceprint) of the second sentence by Michael R. McMullen, CEO of Agilent Technologies Inc., during the earnings conference call on February 14, 2017. The left vertical axis shows the frequency (Hz) of audio signals. The horizontal axis documents the time (seconds), and the right vertical axis (index) and the color represent the audio signals' intensity at each time×frequency location.



pre-trained models such as SpeechVGG for transfer-learning (Beckmann et al., 2019), and economic and finance research also started to adopt the methods (Gorodnichenko et al., 2023; Hu and Ma, 2021; Curti and Kazinnik, 2021; Naik et al., 2016). However, speech analysis based on audio data and DL models in economics and finance is still emerging. In this paper, however, I will look at leveraging the recent advances in AI algorithms for manager speech analysis.

4.3 Deep learning and transfer learning

Now that the audio signals are represented as spectrograms, I can classify them like the other images using neural networks. The model of choice for most image processing tasks is a DL model based on CNNs. The problem with my dataset is that it is relatively small for DL applications. This could lead to overfitting the models to a particular data, which means that the trained models perform well on all the audio signals of manager speech in-sample but would not generalize to other manager speech signals out-of-sample. I address model over-fitting by using dropout layers and transfer learning techniques. Dropout is a regularization technique that can prevent overfitting by randomly dropping out (i.e., set to zero) some of the neurons in a neural network during training. At the same time, transfer learning is a technique in deep learning that involves using pre-trained neural network models as a starting point for training new models on different but related tasks.

Transfer learning can effectively address the problem of overfitting by allowing the

model to leverage knowledge learned from a large and diverse dataset during pre-training. This approach entails deploying a deep learning model proficiently trained on task A involving a substantial dataset to tackle a related task B characterized by a smaller dataset. The advantage lies in the transfer of knowledge acquired during task A training to enhance task B's performance, for example, the hyperparameters and optimal weights of the pre-trained deep learning models. Typically, in transfer learning, the entirety of the pre-trained model is retained, except for the top (classification) layer, which is explicitly retrained on task B. The pre-trained model's output, excluding its classification layer, comprises a high-dimensional vector that appropriately represents various tasks associated with task A, often commonly denoted as a representation.

Considering the relatively modest size of my CEO speeches dataset, I have chosen to implement transfer learning based on pre-trained models and compare results. Because the best choice of model architecture has yet to be conclusive in existing literature, I employ one of the leading pre-trained models for speech representation, namely the SpeechVGG model (Beckmann et al., 2019). Simultaneously, I incorporate a seminal deep convolutional neural network VGG16 architecture for large-scale image recognition and general classification purposes(Simonyan and Zisserman, 2014), serving as a benchmark for comparison. Additionally, I explore the capabilities of the TRILLsson model (Shor et al., 2022), specifically designed to yield a profound representation of speech for "paralinguistic" tasks. The outcomes of the corresponding transfer-learning experiments are currently in the developmental stage (Liu et al., 2023).

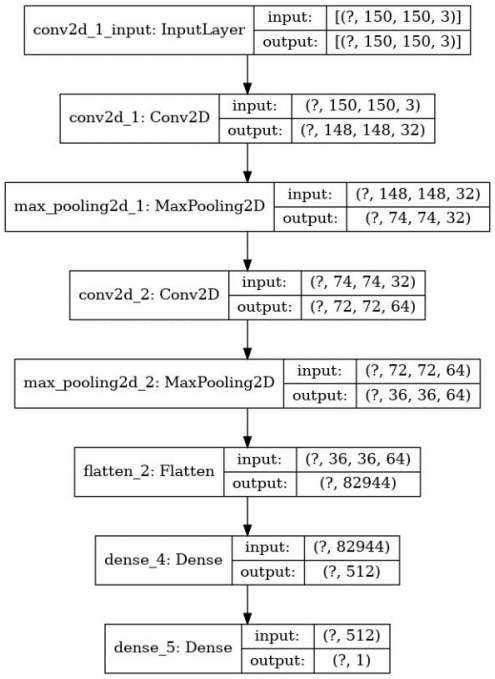
5 Empirical Analysis and Results

This section shows the DL model architectures and model prediction results on firm performance. I trained models with a 1-layer of the CNNs, a 2-layer of CNNs with average-pooling layers, a 2-layer of CNNs with max-pooling layers, a pre-trained VGG16, and a pre-trained SpeechVGG model. I report model architectures and corresponding prediction results in all four models and illustrate a complete training process using a 2-layer CNN model. Figure 5 shows the 2-layer CNN model's model architecture on the left (a) and the model architecture of the transfer-learning model based on VGG16 on the right (b). I report all the other detailed model architectures, including VGG16 and SpeechVGG, in the Online Appendix.

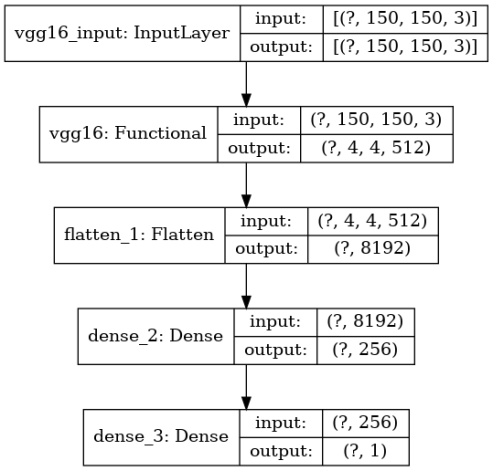
Based on firm performance after the event date of earnings conference calls, I classify

Model Architecture Examples

Figure 5: This figure presents model architectures of deep learning (DL) models. On the left-hand side is a 2-layer CNN model's architecture that includes two CNN layers called "Conv2D" and two max pooling layers. On the right-hand side is a pre-trained VGG16 model (Simonyan and Zisserman, 2014) architecture adopted to binary outputs. The model inputs are the speech spectrograms of managers. The inputs of InputLayer are three-dimensional matrices decomposed from speech spectrograms with a size of $150 \times 150 \times 3$ (heights, widths, and pixels of a graph – red, blue, or green). The model outputs are classification results that are either positive or negative, which is a one-dimensional result. There are 42,487,745 and 16,812,353 parameters for the 2 model architectures, respectively.



(a) A 2-layer CNN model with max-pooling layers



(b) A transfer-learning model based on VGG16

the CEO speeches as either a positive one with a corresponding positive CAR or a negative one that results in a negative CAR. So, the output of the DL models is a binary variable for simplicity as a starting point. I divide the whole sample into training, validation, and final test datasets independent of training processes to train and evaluate models out-of-sample. The training datasets are used to fit the models, while the validation datasets provide an unbiased evolution of a model fit on the training dataset. I decide which model to save during the training process based on the model evaluation, such as prediction accuracy, using the validation datasets. Eventually, after I save the best model from the training and validation process, I use the independent test datasets to evaluate the model performance based on prediction accuracy. Meanwhile, I use mini-batches with a batch

size of 6 spectrograms and an epoch size of 20 times to train models. The batch size is the number of speech spectrograms processed before the model weights are updated based on errors between the model prediction and the realized results. The number of epochs is the number of complete passes through the training dataset. Finally, because the initial model weights and dataset splits are essential for model performance, I validate my models by repeating the out-of-sample prediction 100 times using K-fold cross-validation. I randomize the training, validation, and test dataset splits and the initial model weights every time to ensure the model results' reliability. More specifically, I first fix the order of the datasets, divide the whole sample into four folds ($K = 4$), and take one out of the four as the independent datasets and the rest as training and validation datasets. I repeat this process four times to have four sets of training, validation, and test datasets for a complete round. Finally, I repeated 25 times a complete round to have 100 observations of model prediction accuracy to see the average and confidence intervals of model performances. To my knowledge, I am the first to validate the models instead of focusing on one best-performed model. Next, we can see an example of one complete training process in Figure 6.

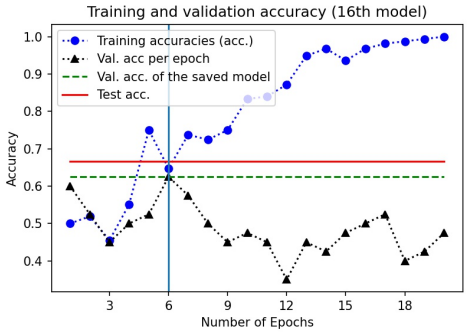
Figure 6a shows that the training accuracies increase over time and reach a prediction accuracy of around 99% on the 20th epoch. Meanwhile, the model is also "learning" in the training datasets over epochs as the validation accuracy increases. The prediction accuracy on the in-sample validation datasets reaches the highest level (around 65%) at the 6th epoch. It starts to decrease until the 12th epoch before the training process ends on the 20th epoch. Finally, I ran a test of the model on the independent out-of-sample test dataset; the model has a prediction accuracy of around 66.7%, which is much higher than a random draw benchmark with an accuracy of 50%. Thus, based on the results shown in Figure 6a, I choose the corresponding model weights from the 6th epoch. In Figure 7b, the out-of-sample prediction accuracy reaches approximately 53.4% with a 99% significant lower bound above 50%. Furthermore, at this stage, the prediction accuracy is solely based on audio data without incorporating textual information and firm characteristics. Therefore, we can interpret the results as economically significant before we move to more stringent tests later in this section.

5.1 Model training and selections

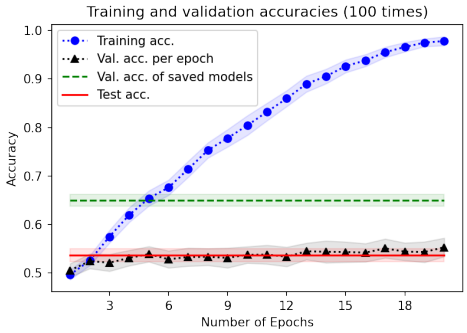
Because the financial analysts are the most direct audiences interacting with the CEO during earnings conference calls and financial analysis are sophisticated finance profession-

An Illustration of the Model Selection

Figure 6: This figure presents the training process and evaluation results on classification tasks based solely on raw audio data and analyst recommendation consensus changes (ΔRECs) of 2-layer CNN models with the max-pooling layers. The dependent variable is binary (negative or positive) based on changes in analyst recommendations one month after the earnings conference call. On the left-hand side is an example of a complete model selection process. The model is trained with a batch size of 6 and an epoch size of 20 times. Training, validation, and test accuracy are presented in the plot using accuracy (%) as the y-axis and the number of epochs as the x-axis. The blue dots represent the model performance per epoch on the training datasets; the black diamonds represent the model performance per epoch on the validation datasets. The horizontal dashed green line shows the model performance on the in-sample validation datasets; the horizontal red line shows the model performance on an independent out-of-sample test dataset; the vertical light-blue lines show the best models' model performance based on validation accuracy in an entire training process (6th epoch in this example). On the right-hand side, the model selection processes are repeated 100 times by randomizing the initial model weights, training, and validation datasets using K-fold cross-validation ($K = 4$). The solid light-blue lines show the 50% threshold (expected accuracy of a random draw) of prediction accuracy. The shaded areas stand for 99% confidence intervals.



(a) one complete model selection process



(b) results of 100 sets of initial model weights

als compared to an average household, I first select the optimal hyperparameters and model weights for different model architectures based on $\Delta RECs$. Then, I employ the selected models to predict other firm performance variables out-of-sample. The main economic mechanisms I am exploring are the market reactions based on an average analyst's perceptions of CEO speeches during earnings conference calls. In this part, I plot the training process and prediction results of short-term firm performance classified by analyst recommendation consensus changes ($\Delta RECs$) in one month after the earnings conference call for the five different model architectures: a 1-layer CNN model (1CNN) with an average-

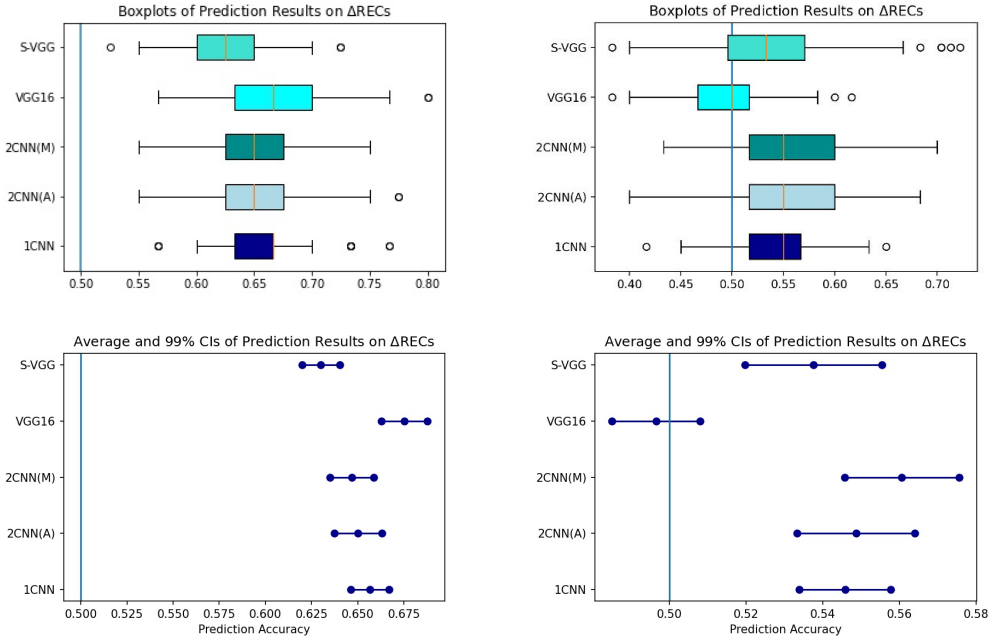
pooling layer[11], a 2-layer CNN model (2CNN(M)) with max-pooling layers[12], a 2-layer CNN model (2CNN(A)) with average-pooling layers, a transfer-learning model based on the pre-trained VGG16 model (VGG16) and a transfer-learning model based on pre-trained SpeechVGG model (S-VGG). The models are trained to predict future firm performance based solely on the second sentence of a manager's speech in an earnings conference call, which contains orthogonal information to textual content and financial information presented by the manager. In total, I have 459 observations due to the database's availability of $\Delta RECs$. I have 359 pairs of speech spectrograms and firm performance in training datasets and 100 pairs in the validation and test datasets. While the positive and negative firm performance observations are equally distributed in the validation and test datasets, I have 197 positive and 162 negative firm performance observations in the training datasets.

[11] In deep learning, max-pooling and average-pooling are common operations used in convolutional neural networks (CNNs) to downsample the spatial dimensions of feature maps. An average-pooling layer takes the average value from the same local region of the input feature map. It computes the average value within each window. The output of the average-pooling layer is a downsampled feature map that is smoother and less sensitive to small variations in the input map.

[12] A max-pooling layer is a pooling operation that takes the maximum value from a local region of the input feature map. The local region is defined by a window or kernel that slides over the input map and selects the maximum value within each window. The output of the max-pooling layer is a downsampled feature map that retains the most important features of the original map while reducing its spatial dimensions.

Model Predictions for Analyst Recommendation Changes

Figure 7: This figure presents the boxplots (top figures) and the average and two-tail 99% confidence intervals (bottom figures) of the chosen models' prediction accuracies of analyst recommendation changes ($\Delta RECs$) on the validation datasets (left) and the independent test datasets (right). The dependent variable is binary (negative or positive) based on changes in analyst recommendations one month after the earnings conference call. The boxplots (sub-figures on top) show minimum without outliers, 25th quantile, median, 75th quantile, and maximum without outliers of prediction accuracies. The void dots are outliers. The two sub-plots on the bottom show the lowest CI boundary (0.05%), average, and highest CI boundary (99.5%) of model prediction accuracies. The solid light-blue lines show the 50% threshold of prediction accuracy. The y-axis shows the model architecture types, and the x-axis shows the prediction accuracy for positive or negative $\Delta RECs$. Along the y-axis, "S-VGG" stands for the transfer-learning model based on SpeechVGG (Simonyan and Zisserman, 2014), "VGG16" stands for the transfer-learning model based on VGG16 (Beckmann et al., 2019), 2CNN(M) stands for my self-trained model with two CNN layers with max-pooling layers, 2CNN(A) stands for my self-trained model with two CNN layers with average-pooling layers, and 1CNN stands for my self-trained model with one CNN layer and a max-pooling layer.



(a) In-sample (b) Out-of-sample

Figure 5 shows prediction accuracies of 100 times model selection processes for five different model architectures on the in-sample validation datasets and the independent out-of-sample test datasets of the chosen models based on the highest value of validation accuracy. The models are trained with a batch size of 6 and an epoch size of 20 times. The training processes are repeated 100 times by randomizing the training and validation datasets and the initial model weights using K-fold cross-validation ($K = 4$). The solid light-blue lines show the 50% threshold of prediction accuracy. In Figure 5a, we can see that all five models perform much better than the chance level in-sample and reach a median accuracy and lower 99% confidence intervals of more than 60%.

Meanwhile, in Figure 5b, we see that the out-of-sample model performances are poorer than the in-sample results. However, four of the five models' median performances and lower 99% confidence intervals are still above the chance level (50%), and three models reach an accuracy of more than 55% on average. Moreover, we see the simpler models with one CNN layer and a max-pooling layer (1CNN), with two CNN layers and max-pooling layers (2CNN (M)), and with two CNN layers with average-pooling layers (2CNN(A)) outperform the more complex architecture - VGG16 and SpeechVGG. While the median performance and the lower 99% confidence interval of SpeechVGG are still reasonably above the chance level, VGG16 shows a problem of model over-fitting.

To further check the out-of-sample model performances, let us check the results on the average accuracies and confidence intervals on the bottom in Figure 5b. As we can tell, the SpeechVGG model's average prediction accuracy is around 54%, and the lowest confidence interval is around 52%, which is above the 50% accuracy threshold. Therefore, I will focus on SpeechVGG for the transfer-learning model for later discussions in the paper. On the other hand, when I further check the simpler models I trained and tuned by myself, we see that the 2CNN(M) model outperforms the other two simpler models - 2CNN(A) and 1CNN. The 2CNN(M) model shows the highest average prediction accuracy, and its lower 99% confidence interval is around 55%, which is well above the 50% threshold. Thus, I will select the 2CNN(M) model as my self-trained model example for later empirical tests. Overall, the preliminary prediction results above show that CEO vocal characteristics are relevant for future firm performance. In the following parts of this paper, I will further test prediction results for other firm performance variables by incorporating textual contents and firm fundamentals based on SpeechVGG and 2CNN(M) models.

5.2 Predicting firm performance

In this part, I summarize the results regarding short-term and long-term future firm performance using multimodal deep learning regression models (MDRMs) that incorporate firm characteristics, the CEO speech's textual features, and the CEO speech's audio features. The models are currently trained to predict long-term firm performance based on the second sentence of a manager's speech, which contains behavioral features orthogonal to textual content and financial information presented by the manager. In this part, I use a simple logistic model to incorporate different features to predict future firm performance. The dependent variables (or labels in the AI literature) - analyst recommendation changes ($\Delta RECs$), unexpected earnings, three-day cumulative abnormal returns (CAR(0,2)), and half-year cumulative abnormal returns (CAR(2,127)) are all binary variables. The dependent variable equals 1 if the value of the corresponding firm performance indicator is above the median value in my sample and equals 0 otherwise. The financial variables are the firm's profitability (earnings per share), past return volatility, book-to-market ratio, size (market capitalization in $ million), and momentum (past half-year cumulative abnormal returns before the earnings conference call (CAR(-127,-2)) following Fama and French (2015) and Ehsani and Linnainmaa (2022). The textual features include the percentage of positive words, percentage of negative words, percentage of uncertainty words, percentage of litigious words, percentage of modal-weak words, percentage of strong modal words, percentage of constraining words, the average number of syllables per word, and average word length following Loughran and McDonald (2011). By using 309 cross-sectional observations[13] in my sample with no missing values in either of the four variables, I show that my deep learning models outperform those that incorporate only financial and textual features on average. I also run robustness checks using other model specifications, such as random forest models and support vector machine models, which are not reported in this version. The main results do not differ from the logistic models.

In the first row of Table 2, we can see that the models that incorporate only financial variables have predictive power for three out of the four firm performance indicators - $\Delta RECs$, CAR(0,2), and CAR(2,127) - on average, but failed to predict the unexpected earnings (UEs). By incorporating the textual features based on Loughran and McDonald (2011), we can see that the prediction results improved in the second row for all four firm performance indicators. Especially regarding UEs, the "Fin.+textual features" model can

[13] I further collected two years of cross-sectional data to construct a panel dataset. However, the results still need to be finalized.

predict the firm performance with an average accuracy of around 56.6% (around 7.5% improvements from the corresponding "Financial fundamentals" model).

I incorporate an audio factor (output from the "dense_5" layer in Figure 3a) generated from my self-trained 2CNN(M) model in the third row. By incorporating a single audio factor, the models outperform the "Fin.+textual features" models regarding analyst-related variables - analyst recommendation changes ($\Delta RECs$) and unexpected earnings (UEs) but failed to explain the market-related variables - CAR(0,2) and CAR(2,127) on average. In the fourth row, when I generate the single audio factor (output from the "dense_3" layer in Figure 3b) from the transfer-learned SpeechVGG model, the models do not outperform the "Fin.+textual features" models in the second row except for predicting the unexpected earnings (UEs) on average. Therefore, we can see that though a single audio factor can be interpreted as vocal sentiment, reducing the audio features to a single audio factor, as in Gorodnichenko et al. (2023), Hu and Ma (2021), and Mayew and Venkatachalam (2012) lose explainable power to the financial fundamentals and textual content of a speech.

Finally, in the last two rows, I incorporate the multi-dimensional audio features output from the second last layers of the deep learning models, as shown in Figure 3 - the "dense_4" layer in Figure 3a and the "dense_2" layer in Figure 3b, in addition to financial fundamentals and textual features. Even though the performance of the multi-dimensional model based on my self-trained 2CNN(M) model only slightly improves from the corresponding single audio factor model in the third row, the prediction accuracies of the multi-dimensional model based on the transfer-learned SpeechVGG model improves significantly. Compared to the "Fin.+textual features" specifications, the "Audio features (SVGG)" models show much better prediction results on analyst recommendation changes (a 14.4% improvement), unexpected earnings (a 6.2% improvement), CAR(0, 2) (a 12.0% improvement), and CAR(2, 127) (a 17.3% improvement). Although these intermediate models are yet to be the best for explaining future firm performance, we can still conclude that the behavioral features of a CEO's speech matter regarding a firm's future performance. Moreover, the models benefit from higher dimensionalities of the unstructured audio data. Thus, it is essential to incorporate unstructured data, like CEO speeches, with granular dimensions for further research. This also distinguishes my paper from Mayew and Venkatachalam (2012) and Hu and Ma (2021) regarding analyzing managers' speeches. Given the flexibility and transparency of my DL models, I can continue the "learning" of my models by incorporating more observations to improve the model weights further and/or tuning hyperparameters of models, such as the number of CNN layers, or exploring more

advanced pre-trained models for transfer learning. Given my DL models' flexibility and transparency, I can continue the "learning" process by incorporating more observations to improve the model weights. Besides, I can fine-tune models' hyperparameters, such as the number of CNN layers, or explore different pre-trained models for transfer learning.

Summary Table of Models' Out-of-sample Prediction Results

Table 2: This table summarizes average out-of-sample prediction results (Pred. Accuracy), corresponding standard errors (Std. Err.), and 99% confidence intervals (CIs) of different logistic model specifications. I use the models to predict firm performance regarding analyst recommendation changes ($\Delta RECs$), unexpected earnings (UEs), three-day cumulative abnormal returns (CAR(0,2)), and half-year cumulative abnormal returns (CAR(2,127)), respectively. The logistic models are trained and tested with an 80-20 sample split based on 309 observations (all four variables are available). The prediction results are based on a 100 times reshuffle of training and testing samples with random initial model weights. "Financial fundamentals" stands for the model that only incorporates financial variables, including the firm's profitability (EPS), past return volatility, book-to-market ratio, size (market cap.), and momentum (CAR(-127,-2)) following Fama and French (2015) and Ehsani and Linnainmaa (2022). "Fin. + textual features" stands for the model incorporating textual features following Loughran and McDonald (2011) in addition to financial features. "+Single audio factor (CNN)" stands for the model incorporating a one-dimensional single audio factor based on my self-trained 2CNN(M) model in addition to the "Fin.+textual features" specifications. The "single audio factor" is the probability that a speech is positive (above median firm performance) with a value range between 0 and 1. "+Single audio factor (SVGG)" stands for the model incorporating a one-dimensional single audio factor based on my transfer-learned SpeechVGG model in addition to the "Fin.+textual features" specifications. "+Audio features (CNN)" stands for the model incorporating multi-dimensional audio features (2nd last layer of the neural network) for constructing the "single audio factor" generated from my self-trained 2CNN(M) model in addition to the "Fin.+textual features" specifications. "+Audio features (SVGG)" stands for the model incorporating multi-dimensional audio features (2nd last layer of the neural network) for constructing the "single audio factor" generated from my transfer-learned SpeechVGG model in addition to the "Fin.+textual features" specifications.

	Prediction results based on logistic models							
	Analyst recommendations changes ($\Delta RECs$)		**Unexpected earnings (UEs)**		**CAR(0,2)**		**CAR(2,127)**	
	Pred. Accuracy (Std. Err.)	[99% CIs]	Pred. Accuracy (Std. Err.)	[99% CIs]	Pred. Accuracy (Std. Err.)	[99% CIs]	Pred. Accuracy (Std. Err.)	[99% CIs]
Financial fundamentals	52.0% (6.3%)	[50.3%, 53.7%]	49.1% (5.6%)	[47.6%, 50.6%]	52.2% (5.9%)	[50.6%, 53.8%]	51.6% (6.1%)	[50.0%, 53.2%]
Fin.+textual features	54.2% (5.3%)	[52.8%, 55.6%]	56.6% (5.8%)	[55.1%, 58.2%]	54.8% (5.4%)	[53.3%, 56.2%]	55.5% (7.0%)	[53.7%, 57.4%]
+Single audio factor (CNN)	54.6% (6.1%)	[53.0%, 56.3%]	56.9% (6.0%)	[55.3%, 58.4%]	53.0% (6.1%)	[51.4%, 54.7%]	54.6% (6.6%)	[52.9%, 56.3%]
+Single audio factor (SVGG)	53.6% (6.6%)	[51.8%, 55.3%]	57.2% (5.7%)	[55.7%, 58.7%]	52.0% (6.0%)	[50.4%, 53.5%]	53.8% (6.0%)	[52.2%, 55.4%]
+Audio features (CNN)	54.6% (5.7%)	[53.1%, 56.1%]	57.1% (6.1%)	[55.5%, 58.7%]	54.3% (5.6%)	[52.9%, 55.8%]	54.4% (6.3%)	[52.7%, 56.0%]
+Audio features (SVGG)	62.0% (8.4%)	[59.8%, 64.3%]	60.1% (6.6%)	[58.3%, 61.8%]	61.4% (8.3%)	[59.2%, 63.6%]	65.1% (7.1%)	[63.2%, 66.9%]

6 Deep Embeddings of CEO Speeches

Following Liu et al. (2023), I try to partially open the "black box" by employing the state-of-the-art transfer-learned deep learning model to extract CEO vocal cues (paralinguistic features). In their study, Liu et al. (2023) demonstrate that the analyst evaluation can be predicted with an accuracy ranging from 60% to 65% based on CEO vocal cues (paralinguistic features) during earnings conference calls using their deep learning model. I first visualize the embeddings of the CEO's speeches to show that emotions cannot fully explain the CEO's vocal cues. This shows an advantage of my approach compared to the existing literature, such as methods used in Gorodnichenko et al. (2023) and Hu and Ma (2021). To improve the interpretability of the acquired embeddings, I developed a vocal emotion classifier utilizing the pre-trained architecture. Subsequently, I compare the identified emotion classes with the classifications related to analyst recommendation changes on the same firm-quarter observation.

6.1 The RAVDESS dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a validated multimodal repository dedicated to emotional speech and song. The speech segment encompasses a spectrum of emotions, including calm, happy, sad, angry, fearful, surprised, and disgusted expressions. Diverse presentation formats are available in the database, including face-and-voice, face-only, and voice-only. The extensive collection comprises 7356 recordings, subject to a thorough evaluation, with each entry rated ten times for emotional validity, intensity, and genuineness (Livingstone and Russo, 2018)[14]. In the subsequent section, I employ the embeddings derived from CEO vocal cues to extract corresponding cues from the RAVDESS datasets. Subsequently, these embeddings are visualized to illustrate the paralinguistic features inherent in the data. Moreover, the vocal emotion classifier

6.2 Examining the deep embeddings for the RAVDESS dataset

Figure 8 illustrates the 1024-dimensional Conformer Applied to Paralinguistics (CAP) model (Liu et al., 2023) mappings using t-distributed stochastic neighbor embedding (t-

[14] All recordings are made freely available under a Creative Commons license and can be downloaded at https://doi.org/10.5281/zenodo.1188976

SNE)[15] following Van der Maaten and Hinton (2008). The three scatter plots are identical; only the colored labels are different. The top plot is labeled according to the seven emotions expressed by the RAVDESS actors. Although some degree of clustering is visible, the class separation is not strong. This indicates that relying solely on emotion does not sufficiently explain the paralinguistic nuances of CEOs. By exclusively employing audio sentiment or emotion labels from CEO speeches, there exists information that remains ignored in the analysis. My approach represents an enhancement of methodologies utilized in previous literature, acknowledging the need for a more comprehensive understanding beyond a singular focus on emotion to capture the intricate dimensions of CEO communication.

In the middle plot, labels are based on the actors responsible for vocalizing the emotions, and the resulting CAP mappings distinctly cluster utterances by the same actor. This observation underscores the significance of individual voice characteristics in explaining the extracted vocal features crucial for predicting firm performance. Moving to the bottom plot, the labels now denote the gender of the actors. Notably, the CAP embeddings for male and female actors exhibit an almost perfect separation, suggesting that the difference in average pitch between males and females is a key factor in explaining the paralinguistic features. Overall, the findings from CAP embeddings underscore the importance of individual characteristics, with gender standing out more prominently than emotion. Once more, this underscores the significance of delving into vocal cues in addition to speech emotion recognition.

[15] t-distributed stochastic neighbor embedding (t-SNE) is a statistical method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map. It is based on Stochastic Neighbor Embedding originally developed by Geoffrey Hinton and Sam Roweis[1], where Laurens van der Maaten proposed the t-distributed variant.[2] It is a nonlinear dimensionality reduction technique for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions. Specifically, it models each high-dimensional object by a two- or three-dimensional point in such a way that nearby points and dissimilar objects model similar objects are modeled by distant points with high probability.

Visualisations of Deep Embeddings

Figure 8: This figure presents visualisations of t-distributed stochastic neighbor embedding (t-SNE) mappings of the raw Conformer Applied to Paralinguistics (CAP) embeddings of the RAVDESS dataset following Liu et al. (2023). The three mappings are identical but labeled according to emotion (top), actor (middle), and sex (bottom). The raw CAP embeddings reveal the actor and especially gender more prominently than emotion.

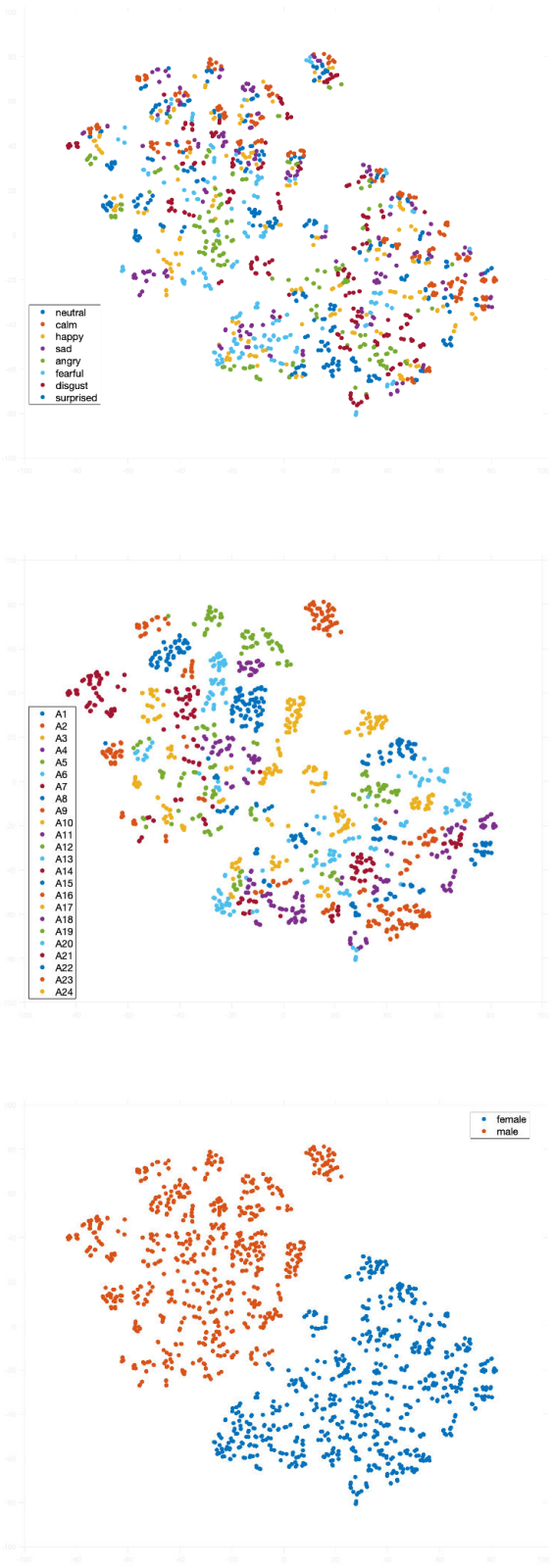
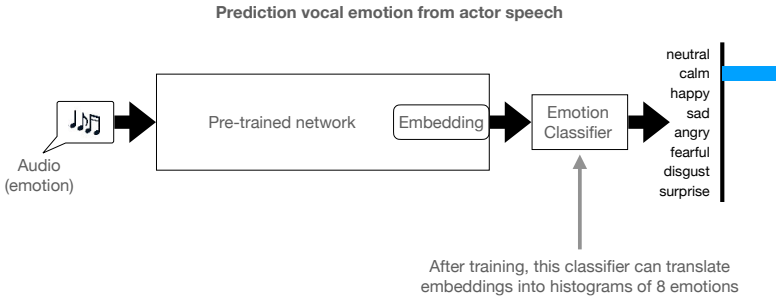


Illustration of Emotion Classifier

Figure 9: This figure shows the utilization of pre-trained deep embeddings based on CEO speeches during quarterly earnings conference calls (QECs). These embeddings serve as input data for training a classifier designed to recognize and categorize the eight distinct emotions outlined in the RAVDESS database following Liu et al. (2023).



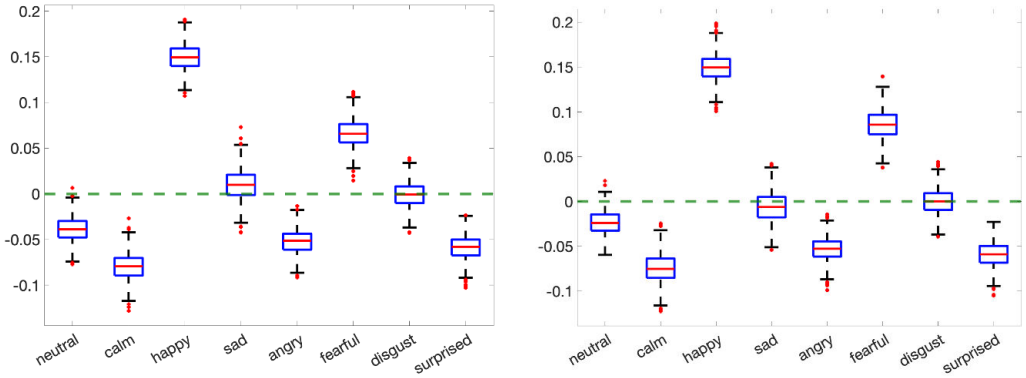
6.3 Mapping the embeddings onto the emotions

To facilitate the interpretability of the deep embeddings, I use the trained classifier on analyst recommendation consensus changes to map the embeddings of the CEO's voices onto the RAVDESS emotions. I employ deep embeddings trained on CEO speeches from quarterly earnings conference calls (CEO QEC) for emotion classification, leveraging the RAVDESS databases. The process involves training a linear logistic model, utilizing the pre-trained deep embeddings, to classify eight RAVDESS emotions. The trained classifier maps each 1024-dimensional embedding onto an 8-dimensional representation, where each element corresponds to a distinct emotion, as shown in Figure 9.

Analyzing the emotional representations of CEO vocal embeddings linked to positive and negative evaluations provides insights into the impact of emotions on the shifts in analyst recommendation consensus changes. In Figure 10, I find that applying the emotion classifier to the CEO-voice embeddings revealed that a positive change in analyst evaluations is reflected in vocal patterns that are more happy and, surprisingly, more fearful.

Emotion Scores: Positive - Negative Embeddings

Figure 10: This figure highlights the variations in emotion scores between CEO speeches classified as positive based on analyst evaluations and those identified as negative speeches following Liu et al. (2023). The y-axis shows the emotion scores, values ranging from 0 to 1, based on the trained emotion classifier based on RAVDESS database. The x-axis shows the eight emotions labeled in the RAVDESS database. Figure 10a on the left-hand side is based on the whole sample, while Figure 10b on the right-hand side is based on the sample with only male CEOs. These figures show that applying the emotion classifier to the CEO-voice embeddings revealed that a positive change in analyst evaluations is reflected in vocal patterns that are more happy and, surprisingly, more fearful.



(a) Male+Female CEOs

(b) All Male CEOs

7 Conclusion and Further Research

To conclude, we can see the predictability of managers' vocal cues to future firm performance. There is evidence that managers' vocal cues are informative about both short-term and long-term firm performance in addition to the quantitative and qualitative contents in quarterly or annual reports or earnings conference calls. I can preliminarily interpret these results as to how the manager says qualitative information, such as business strategies and business plans, are informative, in addition to how the manager summarizes quarterly performances, which are included in quarterly and annual reports and priced by the market. With the most conservative conclusion, vocal cues' or vocal sentiment features predictability supports the social economic, and finance theories that human interactions matter in the information transmission process (Hirshleifer, 2020). Therefore, it is meaningful to move this research further by establishing consistent results and exploring the economic mechanisms behind the economic consequences.

As discussed in Section 5, one technical direction to extend the current research is to improve the DL models by optimizing model weights with more data, fine-tuning hyperparameters, or applying different pre-trained DL models for transfer learning. Furthermore, I can rely on DL visualization to understand the essential vocal features based on model weights, as discussed in Section 6. For example, I can know whether depression detected in the vocal cues is informative in predicting firm performance by following Homsiang et al. (2022).

I have collected two more years of cross-sectional observations to construct a panel dataset to control for firm and time-fixed effects. Meanwhile, to better understand the economic mechanisms and the measures on vocal cues, I will also collect the CEO characteristics from the BoardEx database directly. This approach tests the hypothesis that vocal cues are quality signals that help investors improve their decision-making. For example, if including additional control variables such as CEO education and employment history, the impacts of vocal features drop, then the result supports the hypothesis and vice versa. An investment experiment can also be used to explore behavioral biases. I will follow the structure of (Bohren et al., 2019) and aim to distinguish and quantify inaccurate beliefs versus taste-based channels.

Finally, the methodology and framework in this research can be expanded to other research topics and applications using transfer learning. For example, the analysis of directors' speeches of central banks may predict monetary policies and macroeconomic situations (Gorodnichenko et al., 2023). More related, the economic consequences of speech

analysis and human interactions can be applied to financial results of pension communication, M&As, IPOs, and more qualitative firm perspectives regarding environment, society, and governance (ESG) performance.

References

Baik, B., A. G. Kim, D. S. Kim, and S. Yoon (2023). Managers' vocal delivery and real-time market reactions in earnings calls. *Available at SSRN 4398495*.

Bandiera, O., A. Prat, S. Hansen, and R. Sadun (2020). Ceo behavior and firm performance. *Journal of Political Economy 128*(4), 1325–1369.

Barcellos, L. P. and K. Kadous (2022). Do managers' nonnative accents influence investment decisions? *The Accounting Review 97*(3), 51–75.

Beckmann, P., M. Kegler, H. Saltini, and M. Cernak (2019). Speech-vgg: A deep feature extractor for speech processing. *arXiv preprint arXiv:1910.09909*.

Blankespoor, E., B. E. Hendricks, and G. S. Miller (2017). Perceptions and price: Evidence from ceo presentations at ipo roadshows. *Journal of Accounting Research 55*(2), 275–327.

Bochkay, K., S. V. Brown, A. J. Leone, and J. W. Tucker (2023). Textual analysis in accounting: What's next? *Contemporary accounting research 40*(2), 765–805.

Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2019). Inaccurate statistical discrimination: An identification problem. Technical report, National Bureau of Economic Research.

Burkhardt, F., A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, et al. (2005). A database of german emotional speech. In *Interspeech*, Volume 5, pp. 1517–1520.

Campbell, D. W. and R. Shang (2022). Tone at the bottom: Measuring corporate misconduct risk from the text of employee reviews. *Management Science 68*(9), 7034–7053.

Cen, L., J. Chen, S. Dasgupta, and V. Ragunathan (2021). Do analysts and their employers value access to management? evidence from earnings conference call participation. *Journal of Financial and Quantitative Analysis 56*(3), 745–787.

Choudhury, P., D. Wang, N. A. Carlson, and T. Khanna (2019). Machine learning approaches to facial and text analysis: Discovering ceo oral communication styles. *Strategic Management Journal 40*(11), 1705–1732.

Cohen, L., D. Lou, and C. J. Malloy (2020). Casting conference calls. *Management Science 66*(11), 5015–5039.

Curti, F. and S. Kazinnik (2021). Let's face it: Quantifying the impact of nonverbal communication in fomc press conferences. *Available at SSRN 3782239*.

Custódio, C. and D. Metzger (2014). Financial expert ceos: Ceo s work experience and firm s financial policies. *Journal of Financial Economics 114*(1), 125–154.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic literature 47*(2), 315–72.

DellaVigna, S. and J. M. Pollet (2009). Investor inattention and friday earnings announcements. *The journal of finance 64*(2), 709–749.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee.

Duarte, J., S. Siegel, and L. Young (2012). Trust and credit: The role of appearance in peer-to-peer lending. *The Review of Financial Studies 25*(8), 2455–2484.

Ehsani, S. and J. T. Linnainmaa (2022). Factor momentum and the momentum factor. *The Journal of Finance 77*(3), 1877–1919.

Elliott, W. B., S. M. Grant, and F. D. Hodge (2018). Negative news and investor trust: The role of $ firm and# ceo twitter use. *Journal of Accounting Research 56*(5), 1483–1519.

Engelberg, J. E. and C. A. Parsons (2011). The causal impact of media in financial markets. *the Journal of Finance 66*(1), 67–97.

Ewens, M. (2022). Race and gender in entrepreneurial finance. Technical report, National Bureau of Economic Research.

Ewens, M. and R. R. Townsend (2020). Are early stage investors biased against women? *Journal of Financial Economics 135*(3), 653–677.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of financial economics 116*(1), 1–22.

Francis, B. B., I. Hasan, Y. V. Shen, and Q. Wu (2021). Do activist hedge funds target female ceos? the role of ceo gender in hedge fund activism. *Journal of Financial Economics 141*(1), 372–393.

Garcia, D., X. Hu, and M. Rohrer (2023). The colour of finance words. *Journal of Financial Economics 147*(3), 525–549.

Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as data. *Journal of Economic Literature 57*(3), 535–74.

Giannetti, M., G. Liao, and X. Yu (2015). The brain gain of corporate boards: Evidence from china. *the Journal of Finance 70*(4), 1629–1682.

Gorodnichenko, Y., T. Pham, and O. Talavera (2023). The voice of monetary policy. *American Economic Review 113*(2), 548–84.

Graham, J. R., C. R. Harvey, and M. Puri (2017). A corporate beauty contest. *Management Science 63*(9), 3044–3056.

Heaven, D. (2020). Why faces don't always tell the truth about feelings. *Nature 578*(7796), 502–505.

Hirshleifer, D. (2020). Presidential address: Social transmission bias in economics and finance. *The Journal of Finance 75*(4), 1779–1831.

Homsiang, P., T. Treebupachatsakul, K. Kiatrungrit, and S. Poomrittigul (2022). Classification of depression audio data by deep learning. In *2022 14th Biomedical Engineering International Conference (BMEiCON)*, pp. 1–4. IEEE.

Hu, A. and S. Ma (2021). Persuading investors: A video-based study. Technical report, National Bureau of Economic Research.

Kaplan, S. N., M. M. Klebanov, and M. Sorensen (2012). Which ceo characteristics and abilities matter? *The journal of finance 67*(3), 973–1007.

Kappas, A., U. Hess, and K. R. Scherer (1991). Voice and emotion.

Lacerda, F. (2012). Money talks: The power of voice: A critical review of mayew and venkatachalam's the power of voice: Managerial affective states and future firm performance. *PERILUS*, 1–10.

Liebregts, W., P. Darnihamedani, E. Postma, and M. Atzmueller (2020). The promise of social signal processing for research on decision-making in entrepreneurial contexts. *Small business economics 55*(3), 589–605.

Liu, Z., P. de Goeij, and E. Postma (2023). Predicting company success from ceo speech: A partially explainable deep-learning approach. Technical report, Working Paper.

Livingstone, S. R. and F. A. Russo (2018). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one 13*(5), e0196391.

Long, W. and Y. Zhong (2023). The neglected cohort: The impact of silent majority in social media on stock returns. *Finance Research Letters 52*, 103363.

Loughran, T. and B. McDonald (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance 66*(1), 35–65.

Loughran, T. and B. McDonald (2020). Textual analysis in finance. *Annual Review of Financial Economics 12*, 357–375.

Malmendier, U. and G. Tate (2005). Ceo overconfidence and corporate investment. *The journal of finance 60*(6), 2661–2700.

Mayew, W. J. (2008). Evidence of management discrimination among analysts during earnings conference calls. *Journal of Accounting Research 46*(3), 627–659.

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

Mohamed, A., H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, et al. (2022). Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*.

Momtaz, P. P. (2021). Ceo emotions and firm valuation in initial coin offerings: an artificial emotional intelligence approach. *Strategic Management Journal 42*(3), 558–578.

Mullins, W. and A. Schoar (2016). How do ceos see their roles? management philosophies and styles in family and non-family firms. *Journal of Financial Economics 119*(1), 24–43.

Naik, N., R. Raskar, and C. A. Hidalgo (2016). Cities are physical too: Using computer vision to measure the quality and impact of urban appearance. *American Economic Review 106*(5), 128–32.

Oord, A. v. d., S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Purwins, H., B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath (2019). Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing 13*(2), 206–219.

Qin, Y. and Y. Yang (2019). What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 390–401.

Schuller, B. W. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM 61*(5), 90–99.

Shor, J., A. Jansen, W. Han, D. Park, and Y. Zhang (2022). Universal paralinguistic speech representations using self-supervised conformers. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3169–3173.

Simonyan, K. and A. Zisserman (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Suslava, K. (2021). "stiff business headwinds and uncharted economic waters": The use of euphemisms in earnings conference calls. *Management Science 67*(11), 7184–7213.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance 62*(3), 1139–1168.

Van der Maaten, L. and G. Hinton (2008). Visualizing data using t-sne. *Journal of machine learning research 9*(11).

Wankhade, M., A. C. S. Rao, and C. Kulkarni (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review 55*(7), 5731–5780.

Yosinski, J., J. Clune, A. Nguyen, T. Fuchs, and H. Lipson (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

Zhang, Y., P. Tiňo, A. Leonardis, and K. Tang (2021). A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*.