경영학석사 학위논문

Can You Hear Me?
Vocal Delivery in Earnings Calls and Real-Time Market Reactions

이익발표의 음성 전달력과 실시간 주식시장 반응

2022년 8월

서울대학교 경영대학원

경영학과 회계학전공

김 건 우

Can You Hear Me?
Vocal Delivery in Earnings Calls and Real-Time Market Reactions

이익발표의 음성 전달력과 실시간 주식시장 반응

지도교수 백 복 현

이 논문을 경영학 석사 학위논문으로 제출함

2022년 6월

서울대학교 경영대학원

경영학과 회계학전공

김 건 우

김건우의 경영학 석사 학위논문을 인준함

2022년 6월

위 원 장 황 이 석 (인)

부위원장 신 재 용 (인)

위 원 백 복 현 (인)

Can You Hear Me? Vocal Delivery in Earnings Calls and Real-Time Market Reactions

Gunwoo Kim
College of Business Administration
Graduate School of Business
Seoul National University

Abstract

This study examines the relationship between the vocal delivery of earnings conference calls and real-time market reactions. During earnings conference calls, investors primarily listen to the calls without the written transcripts. Since investors combine non-verbal (audible or vocal) and verbal (readable or textual) information simultaneously when interpreting audible contents (Zahn [1973]), I expect that both verbal and non-verbal cues in earnings calls affect investor decisions. I measure the vocal ambiguity score of each earnings call with a deep learning methodology and find evidence that ambiguous vocal delivery of earnings calls results in a weaker investor response even after controlling for various textual attributes. Furthermore, I find that the effect is less pronounced when a firm conveys bad news and that the effect is more evident when there are more analyst participants in the call. My findings are robust to an alternative measure of vocal delivery and other specifications including *speaker*-level regressions with *executive* and *call* fixed effects. Overall, this study sheds light on a new dimension of information conveyance that affects investor decisions.

Keyword: Earnings conference call, vocal delivery, information conveyance, real-time market reactions, information processing costs
Student Number: 2019-26470

Table of Contents

List of Tables

1 Introduction

This study examines whether vocal components of corporate disclosures affect market reactions. Investors combine verbal (readable contents) and non-verbal (audible vocal cues) information simultaneously when interpreting audible contents (Zahn [1973]). Therefore, I expect that both verbal and non-verbal cues of earnings conference calls (hereafter, earnings calls) are likely to affect investor decisions. In this study, I focus on examining the relationship between vocal delivery, a dimension of non-verbal cues, and market reactions. I apply a deep learning methodology to a large sample of earnings call audio files to calculate the vocal delivery score of managers during earnings calls and find strong evidence that vocal delivery is positively associated with real-time market reactions, even after controlling for various textual attributes. The results indicate that not only the information itself but also how managers convey information to investors matters.

My vocal delivery variable primarily measures the audial clarity of the words spoken by executives during earnings calls. I implement a clarity evaluation metric proposed by Niewiadomski and Akinwale [2015], which matches audio wavelets into their corresponding syllables. My measure jointly captures pronunciation, fluency, and recording quality but is robust to various accents. When vocal delivery is ambiguous, one should expense more cognitive resources to interpret the contents, leading to higher information processing costs. However, high vocal ambiguity does not necessarily mean that the contents are inaudible or impossible to comprehend.[1] In contrast, when vocal delivery is

[1] I manually check the audio files and find very few are physically inaudible.

precise, one could pay less attention to fully understand the contents of a call.[2]

A growing number of disclosures include not only textual components, but also visual or audible components. For instance, 86% of FTSE 350 firms use videos to accompany their corporate disclosures (FRC [2020]), and various communications such as new product presentations and investor/analyst day conferences include multimedia. Moreover, companies provide their earnings calls, shareholder calls, M&A calls, or earnings guidance calls in audio format (S&P [2017]). Such non-verbal aspect of disclosure is important because when people listen to audible information, they cognitively integrate verbal (textual contents) and non-verbal (vocal cues) information to interpret the implication of the contents (Zahn [1973]). Therefore, when investors are exposed to multimedia contents, they simultaneously process verbal and non-verbal cues to form their impression towards the firms (Gundersen and Hopper [1976]) and to make their investment decisions (Barcellos and Kadous [2022]). Nonetheless, the majority of prior studies focus on the textual components of corporate disclosures (for example, see Lee [2016], Brochet, Naranjo, and Yu [2016], Call et al. [2020]) and literature has been relatively silent on the audible dimension of corporate disclosures.

Among a few studies that explore the audible dimension of earnings calls, Mayew and Venkatachalam [2012] find that managerial vocal cues during earnings calls predict future firm performance. Similarly, Hobson, Mayew, and Venkatachalam [2012] identify several verbal cues from earnings calls that precede financial misreporting. More recently, Cao et al. [2020] implement a valence measure to capture the emotion of a speaker. These studies jointly provide

[2]Click on here to listen to an ambiguous vocal delivery sample. Click on here to listen to a clear vocal delivery sample.

evidence that one may infer managers' hidden corporate information from their voice. This study significantly differs from this line of literature in that I focus on another dimension of voice, vocal delivery. The variable of interest captures the audial clarity of the words, rather than emotional affection or cognitive dissonance. Also, I focus on investor response to the differing levels of acoustic delivery, even after controlling for various textual variables.[3]

Earnings call as a medium of disclosure provides an ideal setting to explore the impact of non-verbal cues on investor decisions for several reasons. First, earnings call is a typical voluntary disclosure that 97% of the US firms engage in (Cision [2017]) and a number of investors refer to the calls when making investment decisions.[4] Furthermore, prior studies show that earnings calls deliver well-explained financial information to investors and that there are real-time market reactions in response to earnings calls (Frankel, Johnson, and Skinner [1999], Bowen, Davis, and Matsumoto [2002], Kimbrough [2005], Matsumoto, Pronk, and Roelofsen [2011]). Specifically, they document that earnings calls convey information incrementally useful to accompanying press releases and that investors utilize such information in their real-time decision making (Mayew, Sethuraman, and Venkatachalam [2020]).

Second, audio is the primary non-verbal information channel for earnings calls, compared to other disclosures that include texts or visible contents as their

[3]Similarly, this research question differs from Davila and Guasch [2021], which also studies non-verbal information. They examine whether managers' physical display is informative and whether investors correctly impound that information. On the other hand, this study shows that vocal delivery (non-verbal) affects how investors process verbal information.

[4]In its Shareholder Confidence 365 Study, Cision Ltd. finds that 40 percent of the surveyed individual investors and 65 percent of the surveyed institutional investors refer to earnings calls when making investment decisions. More interestingly, 75 percent of individual investors and 85 percent of institutional investors actually listen to the calls of the stocks that they currently possess.

primary means of communication.[5] Transcripts are made public at least three to four hours after the calls and real-time transcripts are generally not available during the conferences. That is, the investors who make investment decisions during the calls rely on the real-time audios rather than written transcripts. Thus, I may mitigate the confounding effects of other information sources by examining the vocal components of earnings calls.

Last, vocal delivery is less likely to be affected by the complexity of verbal information (Moustroufas and Digalakis [2007], Srikanth, Li, and Salsman [2012]). In other words, even if a specific glossary is complex, one may pronounce it clearly and obtain a high vocal delivery score. Therefore, examining the effect of vocal delivery on market reactions mitigates the effects caused by textual complexity.

I conjecture that vocal delivery in earnings call will affect investor decisions for the following two theoretical backgrounds. First, prior literature illustrates that information conveyance affects investor sentiment and investment decisions (Rennekamp [2012], Bloomfield et al. [2015], Lawrence et al. [2018], Cox, Kreisman, and Dynarski [2020]). In my research design, vocal delivery, which is not econometrically associated with textual complexity, is analogous to the conveyance of information in text disclosures. Therefore, I expect that vocal delivery will affect information processing costs at the margin and investors react differently to varying levels of vocal delivery. Second, auditory phonetics provides evidence that vocal delivery reshapes listeners' impressions towards the speaker (Gundersen and Hopper [1976]). Furthermore, prior liter-

[5] I acknowledge that some calls include presentation slides or videos to accompany conference calls. This additional information might help investors supplement poor vocal delivery. However, presentation slides are typically distributed before the talk, limiting its impact on the real-time market reactions. Also, slides and videos typically do not fully cover what manager says during the call.

ature shows that investor perception affects investment decisions (Blankespoor, Hendricks, and Miller [2017]). In a similar vein, I hypothesize that executive vocal delivery shapes investor perception towards a firm, stiggering incremental market reactions. However, several studies suggest that even if the conveyance of information affects investor perception, it may not affect investment decisions (Agnew and Szykman [2010], Hon-Snir, Kudryavtsev, and Cohen [2012]). Therefore, whether vocal delivery affects investor response remains an open empirical question.

To measure the acoustic delivery of earnings calls, I calculate the vocal ambiguity score (*Vocal_Ambiguity*) for each call using a deep learning algorithm. *Vocal_Ambiguity* captures how "unclear" the audio sounds. I use audio files archived by S&P Global, assuring that the files preserve the audial quality delivered to the audience at the time of earnings calls, without any systematic difference in recording quality. Using all available calls that happen within trading hours for the period of 2008 to 2020 (31,413 calls, 858 GB), I apply E2E (end-to-end) speech-to-text recognition (Wav2Vec) to retrieve texts from audio files. When a machine dictates human voice, it first transforms the audio into visual wavelet fragments and then assigns a syllable to each wavelet fragment. When the wavelet fragment is clean, the machine assigns the corresponding syllable with high certainty score. However, when the fragment is contaminated, the machine assigns a syllable with its best guess. Therefore, the "uncertainty" of syllable assignment approximates vocal ambiguity. The higher the uncertainty is, the more ambiguous a speech is.[6] This methodology evaluates the level of audial clarity as it is perceived by the actual human listeners.

[6] I elaborate this method in further detail in Section 3 and Online Appendix B1 (OB-1).

To find real-time market reactions, I dictate the scripts from audio files and obtain the timestamp for each script line. In earnings call scripts, each line contains a refined script of each executive's speech. That is, when the speaker changes, the script moves to the following line. Therefore, by comparing the machine-dictated scripts with edited scripts provided by S&P, I obtain exactly when each executive starts and ends speaking. This delicate timestamping allows me to examine real-time market reactions in response to executive speech.[7]

I use vocal ambiguity scores to measure the acoustic delivery of earnings calls. I separately obtain vocal ambiguity scores for presentation and discussion sessions of each call. Then I relate vocal ambiguity scores to real-time market reaction variables (abnormal volume and abnormal returns).[8] The results provide a strong support for the hypothesis that the vocal presentation of information influences real-time investor responses. Even after controlling for other factors that are known to affect market reactions, I find a negative association between vocal ambiguity scores and market reaction variables. In terms of economic significance, a one standard deviation increase in vocal ambiguity is associated with a 2.36% (3.66%) decrease in abnormal trading volume and a 2.69% (3.42%) decrease in abnormal absolute returns during the presentation (discussion) session. The relation is robust under *speaker*-level regressions with *executive* and *call* fixed effects. The results provide compelling evidence that the audible environment affects concurrent investment decisions. When speaker mumbles, investors have difficulty processing the information immediately, lead-

[7] See Online Appendix A (OA-A) that details how I obtain the timestamps.

[8] For example, if the presentation session lasts for 22 minutes and 15 seconds in a call that starts from 14:00, I use abnormal trading volume and return between 14:00:00 and 14:22:15.

ing to a smaller abnormal trading volume and lower abnormal absolute returns.[9]

To examine the cross-sectional variation in the relation between vocal delivery and real-time market reactions, I further investigate whether the types of information and the number of participants in earnings calls play a moderating role. I show that the vocal effect is more pronounced when a firm conveys good news rather than bad news. This evidence implies that investors pay more attention when interpreting negative information (Soroka [2006], Wu et al. [2011]) and they tend to focus more on the contents of the information rather than how clearly it is delivered. Furthermore, I find that the earnings call audio has more impact on real-time investor decisions when there are more analyst participants during the discussion session. This finding is consistent with the view that the number of analyst participants approximates the real-time demand of earnings call information and affects the magnitude of real-time market reactions (Hobson, Mayew, and Venkatachalam [2012], Brochet, Naranjo, and Yu [2016]). Overall, my findings suggest an important role of information conveyance during earnings calls.

This paper makes several important contributions to disclosure literature. First, this study provides novel evidence that non-verbal information environment is associated with real-time market reactions. Even though the audible dimension constitutes a critical part in investor decision making during earnings calls, literature has been relatively silent on this issue. This paper presents a large-sample evidence that vocal environment is associated with investor decision and that acoustically ambiguous speech leads to a lower magnitude of market reactions. This finding fills a void in the literature regarding investors'

[9] Note that high vocal ambiguity does not necessarily mean that the information is inaudible. Inaudible contents represent an extreme end of vocal ambiguity. Generally, high vocal ambiguity samples are "hard to understand," rather than "impossible to understand."

immediate response to audible contents. Practically, it articulates that firms need to pay attention to the vocal delivery to promote more salient market reactions during the earnings calls.

Second, this research adds to the line of literature that examines whether the conveyance of information matters. Extant research explores the effect of information conveyance on investor perception. For instance, prior studies generally concur that linguistically complex contents require more cognitive resources to interpret them, leading to higher information processing costs for investors (Rennekamp [2012], Blankespoor, deHaan, and Marinovic [2020]). However, as Bushee, Gow, and Taylor [2018] point out, it is difficult to differentiate linguistic complexity from contents complexity in empirical setting. This is because textual complexity measures such as Fog or Bog index capture both linguistic and contents complexity. Nonetheless, my measure of vocal delivery is not econometrically associated with Fog index and erroneous grammar usage, providing an ideal setting to clearly demonstrate the effect of information conveyance on investor reaction.

Third, empirically, I provide a more precise method to measure real-time market reactions during earnings calls in a large sample. By using audio timestamping algorithm, I match each line of transcript with exact real-time market reaction variables. Several prior studies use trade-and-quotes (TAQ) data to obtain proxies for concurrent market reactions (Hollander, Pronk, and Roelofsen [2010], Matsumoto, Pronk, and Roelofsen [2011], Brochet, Naranjo, and Yu [2016]). However, they approximate the time when the presentation session ends with the average number of words spoken per minute. Also, they fail to obtain detailed timestamps for each line of transcript. One exception

is Mayew, Sethuraman, and Venkatachalam [2020] where they manually listen to 2,455 calls to assign timestamps to each turns-at-talk, which is obviously labor-intensive. Although they acknowledge that automating this process is not feasible, my deep-learning methodology enables the large sample timestamping (e.g., 31,224 calls in this study). I believe the introduction of this automated algorithm, which allows a large sample analysis of real-time market reactions during earnings calls, would open up ample research opportunities that were previously deemed infeasible.

Lastly, to the best of my knowledge, this research is the first to utilize a large sample of earnings call audios to analyze their vocal characteristics. Previous literature on vocal characteristics analyzes a limited sample and cannot be easily generalized. This is because audio analysis is highly time- and resource-consuming due to its high computational costs. However, I implement a deep-learning methodology (Wav2Vec) to explore the complete big data set of earnings call audios. I examine all 31,224 earnings call audios that happen within trading hours of the US-listed firms from 2008 to 2020. This approach not only increases the power of my econometric tests, but also makes my results more generalizable.

The remainder of the paper is organized as follows. In Section 2, I review prior literature and develop the hypotheses. Section 3 details empirical research design and Section 4 reports the results. Finally, Section 5 concludes.

2 Literature Review and Hypothesis Development

2.1 Earnings Calls and Their Economic Consequences

Prior literature focuses on whether earnings calls are informative and whether investors react to the calls. Tasker [1998] suggests that conference calls supplement financial statements by providing incremental information content. Frankel, Johnson, and Skinner [1999] find an increase in abnormal trading volume during earnings calls, and Bowen, Davis, and Matsumoto [2002] show that earnings calls improve analyst forecast accuracy. Kimbrough [2005] also finds that earnings calls reduce the magnitude of post-earnings announcement drift, which represents investor underreaction. Matsumoto, Pronk, and Roelofsen [2011] divide the earnings calls into presentation and discussion sections by counting the average number of words spoken in a minute and then compare the relative informativeness of each section. They demonstrate that both periods are additionally informative to earnings press releases and that the discussion period generally is more informative than the presentation period.

Provided that earnings calls generate immediate market reactions, several studies aim at identifying textual characteristics of earnings calls that affect investor decisions. Davis et al. [2015] find a manager-specific tone in earnings calls by calculating the unexplained portion in the textual sentiment of earnings call scripts. They portray that managers' background may explain manager-specific tones and that investors react in response to them. Similarly, Price et al. [2012] show that earnings call scripts' positive or negative textual tone is associated with future stock returns. Brochet, Naranjo, and Yu [2016] show the use of non-plain English and erroneous expressions incur lower market reactions

during the call. Lee [2016] finds that investors discount the lack of spontaneity in managers' speech during the discussion period. Similarly, Call et al. [2020] show that managers' use of humor during the discussion period leads to more favorable media coverage and stock recommendation revisions. Taken together, prior studies agree that textual contents of earnings calls affect the capital market and investor behavior. However, research has been relatively silent on vocal components of earnings calls.

Among limited empirical evidence on non-verbal cues of earnings calls, Mayew and Venkatachalam [2012] analyze when analysts scrutinize managers during the discussion period. They then measure the positive or negative affection from audio files recorded during the calls. They find that positive or negative emotions during the calls predict firms' future financial performance. In a similar vein, Hobson, Mayew, and Venkatachalam [2012] also explore the vocal dissonance markers in CEO speeches. Specifically, they show that audial dissonance markers are associated with a higher likelihood of reporting irregularities. Both studies suggest that vocal cues observed during earnings calls, in addition to speech composition, have additional information regarding the future behavior of firms. However, due to high computational costs, they only analyze earnings call audios of 2007, making it difficult to draw generalizable inferences. Furthermore, since textual contents may influence affection and vocal dissonance (Zuckerman, DePaulo, and Rosenthal [1981], Vrij [2008]), one may not completely rule out the possibility that the results partially stem from verbal rather than vocal components of earnings calls. These two studies underscore that managers' voice conveys information hidden in textual contents.

2.2 Information Conveyance

Several experimental studies illustrate that how firms convey information, in addition to the contents of the information, affects investor behavior and perception. Rennekamp [2012] finds in her experiment that more readable corporate disclosure leads to stronger investor reaction. She relates her results to processing fluency theory in psychology and portrays that the investors who find the disclosure easier to understand are more likely to rely on it. Huang and Liu [2007] and Kacperczyk, Van Nieuwerburgh, and Veldkamp [2016] attribute this shift in attention to the rational resource allocation of investors in that the investors are likely to assign higher importance to tasks that they consider easier to process. Asay, Libby, and Rennekamp [2018] show that personal pronoun usage is positively associated with the perceived credibility of a speaker. It shows that information display reshapes investors' perception of the firms and that the investors may change their decisions.

As Blankespoor, deHaan, and Marinovic [2020] highlight, investors may place more weight on the information that is easy to understand and that grabs more attention, leading to lower information processing costs. The presentation of information has the potential to incur variations in market reactions. For example, Lawrence et al. [2018] find that earnings announcement promotions in Yahoo Finance lead to high abnormal returns on the earnings announcement date. Since the promoted firms are randomly selected, they infer that the visibility of earnings news affects investor attention. In addition, Cox, Kreisman, and Dynarski [2020] experiment to find that the visualization of textual disclosure helps investors better understand the contents and ultimately reduce investment fees. However, in contrast, Beshears et al. [2011] find that even

though the summary of financial information does affect investor perception, it does not affect portfolio construction.

2.3 Vocal Delivery and Its Effects

In auditory phonetics, prior research focuses on the effect of speech delivery on listeners. Gundersen and Hopper [1976] perform experiments to examine the effect of speech delivery (vocal dimension) and speech composition (verbal dimension) on speech effectiveness. This study finds that when the audience listens to a speech, they tend to change their perception of the presenter. Specifically, their study shows that when the audience listens to a presentation with superior delivery, they tend to perceive the presenter as professional and trustworthy. Furthermore, the study suggests that verbal and non-verbal dimensions jointly determine overall speech efficiency.

Several studies investigate the relation between vocal delivery and listeners' cognitive burden. Van Engen, Chandrasekaran, and Smiljanic [2012] and Ward et al. [2016] show that clear speech improves the short-term memory and recognition of spoken phrases for listeners. Similarly, Van Engen and Peelle [2014] demonstrate that when the audience listens to easy-to-understand audio, the listeners need less cognitive effort to process the information. Specifically, they expect cognitive mismatches between their vocal expectations and input information to occur when listeners are exposed to non-clear audios. As anticipated, they show that the listeners put more effort into comprehending heavily-accented speech in an experiment.

Recently, Barcellos and Kadous [2022] explore investor reactions to earnings calls in a controlled experiment. In their experiment, they change the

accent of the same earnings call transcript and let the participants listen to the two versions of the same earnings call. They then focus on the reconciliation of conflicting impressions regarding CEOs. They argue that investors' stereotypes conflict with each other when CEOs have a non-native accent, leading to increased processing efforts. Since bad news requires more thorough cognitive processing, investors are likely to form more positive impressions of the CEOs and react less negatively in response to bad news.

2.4 Hypothesis Development

Summarizing the discussion above, earnings calls convey information that is additionally useful to the information released in accompanying press releases and that investors react immediately to such information. Earnings calls are audible, and investors do not have access to refined transcripts while listening to them. Since humans interpret both verbal and non-verbal components simultaneously while listening to audible contents (Zahn [1973]), the vocal (or non-verbal) characteristics of earnings calls are likely to affect investor decisions. So far, research has focused on exploring the verbal dimension of earnings calls and examined how verbal features affect investor behavior. Textual characteristics measured from transcripts such as grammar (Brochet, Naranjo, and Yu [2016]), spontaneity (Lee [2016]), and humor (Call et al. [2020]) influence investor behavior during or right after the conferences. However, research has generally been silent on exploring the effect of earnings calls' audible dimension on investment decisions.

I conjecture that earnings call delivery may affect investor decisions for at least the following two channels. First, vocal delivery can affect information

processing costs at the margin. Prior literature finds that how firms convey information does matter. The summary (Bloomfield et al. [2015]) and readability (Rennekamp [2012]) of financial statements, webpage promotion (Lawrence et al. [2018]), and visualization of complex information (Cox, Kreisman, and Dynarski [2020]) are positively associated with increased investor attention and reaction. Even though a firm conveys the same information, how the firm conveys such information affects investor response. Similarly, when executives deliver a clearer speech, investors find the information easier to understand. Shah and Oppenheimer [2007] find that investors are likely to place a higher weight on simple information when making judgments. Therefore, I expect positive market reactions during earnings calls with superior delivery.

Second, vocal delivery can affect investors' perception of the speaker. Auditory phonetics finds that listeners are likely to change their perception of the speaker to be more trustworthy and professional after listening to a speech with good delivery (Gundersen and Hopper [1976]). Similarly, investors listening to earnings calls are likely to change their perception of executives depending on the acoustic delivery of the calls. Blankespoor, Hendricks, and Miller [2017] show that investor perception towards CEOs is positively associated with IPO pricing. Huang, Snellman, and Vermaelen [2020] find that perceived executive trustworthiness is positively related to the long-term excess return. In sum, prior literature agrees that investor perception impacts financial investment decisions. Therefore, I posit that investors who perceive executives as more credible while listening to earnings calls make more active investment decisions.

However, it is also possible that the audial conveyance of information does

not affect investor decisions. Simon [1990] provides a theoretical setting of the bounded rationality model, which asserts that investors take their best option considering all the restrictions they face. In an empirical setting, Cohen and Kudryavtsev [2012] find that only investor behaviors based on informed decisions impact the market and that heuristics are canceled out in the stock market as a whole. In this research, the vocal delivery is rather a conveyance of information than the contents. Therefore, if investors react only to the contents, they will place less weight on vocal cues when making investment decisions.

The discussion above yields the first hypothesis in an alternative form:

H1: *Ceteris paribus*, earnings calls with superior delivery will lead to more active market reactions during the conferences.

To deepen our understanding regarding the mechanism of **H1**, I perform two cross-sectional analyses. First, Soroka [2006] finds that bad news has a larger impact on investor sentiment and investment decisions. His argument is in line with that of Weiner [1985] and Mercer [2005] that people think more deeply about the causes of negative corporate consequences than they do about positive consequences. As investors expense more cognitive resources on the contents when interpreting negative news than when interpreting positive news, the influence of vocal delivery could decrease for the bad news. Thus, I hypothesize that vocal delivery will have less impact on market reactions when managers convey negative information. I formally state **H2a** as follows:

H2a: *Ceteris paribus*, the relation between the vocal delivery of earnings calls and market reactions, if any, will be less pronounced when executives convey more negative news.

On the other hand, I also examine the demand side of earnings call informa-

tion. I conjecture that when more investors listen to the calls, more investors are likely to be affected by vocal delivery, leading to a more pronounced vocal delivery effect. Prior literature finds that the number of analyst participants during earnings calls has a significant positive impact on market reactions (Hollander, Pronk, and Roelofsen [2010], Brochet, Naranjo, and Yu [2016]). Therefore, I infer that earnings call information is better disseminated and incorporated into the financial market when there are more participants during the calls. Since the exact number of listeners during each earnings call is not available, I use the number of participants during discussion sessions to proxy for the demand for earnings call information. Hence, when there are more analyst participants, I expect the magnitude of vocal delivery effect on market reactions to be more salient. I formally state **H2b** as follows:

H2b: *Ceteris paribus*, the relation between vocal delivery of earnings calls and market reactions, if any, will be more pronounced when there are more analyst participants during earnings calls.

3 Data and Model

3.1 Data

I examine a large sample of earnings call transcripts and audio data of U.S. firms from 2008 to 2020 obtained from S&P Global. First, I download all available transcripts of 119,211 distinct earnings calls. Then, I match each script with its corresponding audio files. I eliminate 1,165 calls with missing audio files or erroneously truncated audios. Next, since the primary purpose of this study is to measure the real-time market reactions during the calls, I drop 86,633 calls that happen when the stock market is closed. Then, I delete short

earnings calls following Brochet, Naranjo, and Yu [2016]. Specifically, I drop 5,644 calls shorter than 21.78 minutes, which correspond to 1% of the remaining calls.[10] I require financial data from Compustat and CRSP and M&A data from SDC Platinum. This process leads to 30,224 distinct firm-quarter call observations from 1,946 firms.[11] Furthermore, I require real-time millisecond market data from trade and quote (TAQ) as well. I drop several observations in each regression due to data unavailability in TAQ.

In script-line level regressions, I analyze the scripts during the presentation sessions line-by-line. Each line corresponds to one speaker. I start with 7,031,038 lines and drop lines without valid audio files (81,526 lines) and lines that are broadcasted when the stock market is closed (4,941,953 lines). Also, I restrict the observations to presentation sessions of earnings calls for the following reasons. First, the average audio length of each script line spoken by executives is much shorter in discussion sessions (28.78 seconds) than in presentation sessions (314.71 seconds). Therefore, even though investors react almost immediately in response to vocal cues, their reactions may happen after the speech ends during discussion sessions. In contrast, during presentation sessions, it is far less likely that a significant portion of real-time investor reactions will happen after the speech ends. Second, my vocal delivery measure could be biased when assessing relatively short audio fragments. Since the vocal ambiguity score is an averaged self-entropy, several outliers could largely affect

[10] This is because short calls do not follow the typical composition (presentation and discussion) of earnings calls. Some of them skip discussion sessions, while others integrate presentations and discussions.

[11] My sample reconciliation is similar to Matsumoto, Pronk, and Roelofsen [2011], and their sample consists of 10,062 firm-quarter observations during a three-year period (from 2003 to 2005). My sample spans from 2008 to 2020 (a 13-year period) but I require valid audio files and delete short calls. Considering two more restrictions that I impose on the sample selection procedures, the number of my final sample is comparable to that of Matsumoto, Pronk, and Roelofsen [2011].

the results, especially when the input data is relatively small.[12] By restricting the sample to presentation sessions, I drop 1,814,023 lines. Also, I require each line to be longer than 15 seconds[13] and to be delivered by an executive, not by an operator. This procedure yields 107,316 observations or 3.55 lines per call. Table 1 illustrates the summarized sample selection procedure.

3.2 Timestamps and Real-Time Market Reactions

To examine the real-time market reactions in response to differing vocal deliveries during earnings calls, I calculate the exact start and end time of each script line, labeled as "timestamp."

I briefly discuss the methodology to obtain timestamps for all script lines and leave the detailed explanation in Online Appendix A (OA-A). I first convert audio files to have exactly 16,000 frames per second and transform the audio information into number arrays representing the audio data wave functions. Then I use Wav2Vec2.0, an E2E (End-to-End) deep-learning-based speech recognition algorithm, to convert the number arrays into corresponding syllables. I merge the retrieved syllables to obtain the machine-created transcripts. Next, I divide the retrieved texts into input groups containing 50,000 frames. Then, I compare the text similarity between each input group and edited transcripts from S&P Global using the 4-gram metrics. This process matches each input group to its corresponding script line. Since the converted audio files have 16,000 frames per second, each input group is 3.125 seconds long. Therefore, I

[12] Even though I acknowledge that script-line level observations could be noisy during discussion sessions, I also report script-line level regression results during discussion sessions in Section 4.

[13] Lines less than 15 seconds are less likely to convey new information. Furthermore, my timestamps allow a maximum of 3.125-second error for each line. 15-second cutoff corresponds to the 1% of the total observations.

deduct the timestamp for each script line by adding the number of input groups matched to each line. The timestamp is precise, with an error margin of only 3 seconds.

After obtaining the timestamp for each script line, I obtain the corresponding real-time market reactions. Since I know exactly when each speech commences and ends, I calculate the abnormal trading volume and abnormal absolute returns for each script line. Furthermore, to spot the time the discussion session begins, I flag the first inquiry from the analyst of each earnings call. Previous literature approximates the length of the earnings call and presentation session by dividing the number of words by 160, which is the average number of words that people speak in one minute (Matsumoto, Pronk, and Roelofsen [2011], Brochet, Naranjo, and Yu [2016]).

3.3 Vocal Ambiguity Measure

To capture the vocal delivery of each script-line, I follow the Goodness of Pronunciation (GoP) developed by Witt and Young [1997]. By computing self-entropy (Zou et al. [2018], Saporta et al. [2020]), I extract a portion of their measure to compute vocal ambiguity. I briefly outline the calculation here and leave the details in Online Appendix B1 (OA-B1).

When a machine translates audio files into texts, it goes through a two-step conversion. First, it transforms the audio into a wave function, represented by an array of numbers. Then, Wav2Vec2.0 decodes the array of numbers and assigns it the corresponding syllable. The classifier contains 32 syllables and assigns each array to one of the syllable categories. Ideally, if an array is clear (or easily classified), the machine assigns one syllable with certainty to the ar-

ray. However, in real data, number arrays are unclear and cannot be assigned with 100% certainty. In such a case, Wav2Vec2.0 assigns a probability vector with 32 components instead of assigning one single syllable to an array. The vector contains "respective logistic probability" that the array is assigned to each one of the syllables. For instance, if an array can be assigned to category 1 with 100% certainty, the vector would be $(1, 0, 0, \cdots, 0)$. However, if an array resembles all categories to some extent, the vector would be $(p_1, p_2, \cdots, p_{32})$, where $p_i > 0$.

I calculate the measure of vocal ambiguity with the logistic probability vector. The variable of interest (*Vocal_Ambiguity*) is the self-entropy of the logistic probability vector (Equation OA-(1)). In the extreme case where the audio data is assigned sole syllable with 100% certainty, *Vocal_Ambiguity* takes the value of zero. However, as the logistic probability distribution becomes more widespread (i.e., the audio is unclear so that the machine cannot assign an appropriate syllable), *Vocal_Ambiguity* increases.

I separately measure vocal ambiguity scores for presentation and discussion sessions of each earnings call. In this measure, even native speakers may score low. This evaluation aims to assess whether a speaker has clear or ambiguous vocal delivery, not whether a speaker has a native accent. I acknowledge that even though the speaker's pronunciation mainly affects the score (Baevski et al. [2020]), this measurement method is subject to other confounding factors such as recording environment or noises. However, I do not attempt to control for noise since the audio files are the actual files that investors listen to during the conferences. Therefore, my measure of vocal ambiguity collectively considers the environment of information conveyance.

3.4 Regression Design

Since I aim to examine the effect of vocal delivery on real-time market reactions, I obtain real-time market trading data from TAQ, following Matsumoto, Pronk, and Roelofsen [2011] and Brochet, Naranjo, and Yu [2016]. Even though one-day abnormal trading volume and abnormal absolute return are commonly used to investigate the concurrent market reactions, those measures are likely to be affected by factors other than vocal delivery. Therefore, while testing the speaker-level effect, I use the exact timestamps and calculate the precise corresponding market variables.

To test **H1**, I separately estimate the ordinary least squares regressions for the presentation and discussion sessions as follows.

$$Volume(Return)_{it}^{PPT(Q\&A)} = \beta_1 Vocal\_Ambiguity_{it}^{PPT(Q\&A)} + \omega Z_{it} + \tau S_{it} + FE + \epsilon_{it} \tag{1}$$

Here, $Volume_{it}^{PPT(Q\&A)}$ refers to the abnormal trading volume of firm i during the presentation (discussion) session of quarter t's earnings call. For instance, for a call of firm i starts at 14:00 PM on September 26, 2020 and lasts for the following 58 minutes 35 seconds, I first calculate that the discussion session commences exactly 23 minutes 15 seconds after the call starts. Then, I obtain the abnormal trading volume from 14:00:00 to 14:23:15 on September 26, 2021 as $Volume_{it}^{PPT}$ and the abnormal trading volume from 14:23:15 to 14:58:35 as $Volume_{it}^{Q\&A}$. Similarly, for $Return_{it}^{PPT(Q\&A)}$, I calculate the abnormal absolute return for the exact time period for each presentation and discussion session. I follow Brochet, Naranjo, and Yu [2016] to calculate abnormal trading volume

and absolute returns. Specifically, to calculate trading volume, I take the natural logarithm of the sum of the total number of shares traded during a specific period of time. To obtain an absolute return, I scale the absolute difference between the starting and ending quotes by the starting quote. Then, I subtract the median trading volume and absolute returns from the same time period of the same day over the preceding two weeks from the earnings call.

The variable of interest is $Vocal\_Ambiguity_{it}^{PPT(Q\&A)}$, measured at the presentation (discussion) level. Using the script-line level *Vocal_Ambiguity* scores discussed in the previous section, I construct the presentation (discussion) level variable by taking the weighted average of the scores. I use *Audio_Length* of each script line as weights. Specifically, I obtain $Vocal\_Ambiguity_{it}^{PPT(Q\&A)}$ using the following Equation (2):

$$Vocal\_Ambiguity_{it}^{PPT(Q\&A)} = \frac{\sum_q (Audio\_Length_{qit} \times Vocal\_Ambiguity_{qit})}{Audio\_Length_{it}} \tag{2}$$

Here, q denotes the line included in the presentation (discussion) session of firm i's quarter t earnings call. While calculating the weighted average vocal ambiguity score, I exclude script lines that are less than 15 seconds[14] since short lines are generally operator instructions, short answers, or greetings. Therefore, higher $Vocal\_Ambiguity_{it}^{PPT(Q\&A)}$ indicates bad vocal delivery. I expect β_1 to be negative and statistically significant.

Next, I include a variety of firm-level and script-level variables. Here, Z_{it} denotes a vector of firm-level control variables of firm i in quarter t and ω is the corresponding coefficient vector. Following prior literature, I include the size,

[14]Short lines account for 0.99% of the total observations in the data.

return on asset, market-to-book ratio, leverage ratio, earnings volatility over the preceding five quarters, loss indicator, the number of segments, M&A indicator, absolute earnings surprise, volatility of the stock returns during the preceding two quarters, positive return indicator for the next quarter, an indicator that equals one if a call is held on Friday, the number of analyst participants during the discussion session of a call, the time period between the fiscal quarter-end and earnings call date, the fourth quarter indicator, and the stock market return of one day before the earnings call. Also, I include $Vocal\_Ambiguity_{it}^{PPT}$ when regressing $Vocal\_Ambiguity_{it}^{Q\&A}$ on $Volume(Return)_{it}^{Q\&A}$ to mitigate the concern that the abnormal volume (return) of presentation session heavily affects the market reactions of the following discussion session.

Then, I include script-level variables calculated at the presentation or discussion session-level, respectively.[15] S_{it} denotes the vector of script-level control variables, and τ is the corresponding coefficient vector. I include the presentation (discussion) level Fog index, the length of each session in minutes, and the linguistic tone measured with the financial dictionaries provided by Loughran and McDonald [2011]. Specifically, I subtract the number of negative words from the number of positive words and scale it with the number of total words in each session. Furthermore, I replicate the grammar measure of Brochet, Naranjo, and Yu [2016] to control for the effect of incorrect grammar usage on real-time market reactions. I calculate (*i*) the number of grammatical errors other than punctuation errors, scaled by the number of total words, (*ii*) the number of passive voice usages scaled by the number of total sentences, and (*iii*) the number of abnormal usages of the article "the". Then, I standardize

[15] The scripts contain some instructions in brackets [ ]. For instance, they mark utterances with [ph] and skip operator instructions with [Operator Instructions]. Therefore, to obtain cleaner textual information, I parse out fragments that are within the brackets [ ].

the three numbers and take the average to obtain $Grammar\_Error_{it}$. Other than control variables, I include year and industry fixed effects to mitigate effects from omitted variables. I detail the variable descriptions in Appendix A.

To examine whether the vocal delivery effect persists in script line-level, I estimate the following Equation (3):

$$Volume(Return)_{qit}^{PPT(Q\&A)} = \beta_1 Vocal\_Ambiguity_{qit}^{PPT(Q\&A)} + \tau S_{qit} + FE + \epsilon_{it} \tag{3}$$

In a single call, an average of three executives take turns speaking during the presentation session, and their vocal deliveries differ. Therefore, I conjecture that investors react differently to executives with varying vocal deliveries, even within a single call. The regression specification is similar to Equation (1), except that I use script-line level variables. The subscript q denotes the line of speech delivered by an executive during the quarter t earnings call of firm i. Similarly, $Volume(Return)_{qit}^{PPT}$ corresponds exactly to the time when line q of the script is being broadcasted to investors. S_{qit} includes the Fog index, audio length, grammar error, and net tonality of each script line q. Also, I include *executive* fixed effects and *call* fixed effects to control for other omitted variables.[16]

[16] In an untabulated test, I include executive-level control variables (E_{qit}) that are known to affect vocal delivery. First, since native speakers are likely to exhibit clearer delivery, I include an indicator variable that equals one if an executive is from an English-speaking country like the United States, Canada, United Kingdom, or Australia. This does not imply that all natives have superior delivery. As illustrated in the previous section, non-natives can score high in my evaluation metric if they speak clearly, even in a non-native accent. To consider the gender and age differences in delivery, I include the age and gender of each executive as control variables. This specification yields qualitatively similar results even with a reduced sample.

3.5 Descriptive Statistics

Table 2 presents the descriptive statistics for the variables I include in regression analyses. In Panel A, I report summary statistics for presentation and discussion level variables. The mean values of abnormal absolute return during presentations and discussions are 0.005 and 0.006, respectively. These figures are comparable to the abnormal absolute returns obtained in Matsumoto, Pronk, and Roelofsen [2011]. The mean values of abnormal trading volume during presentations and discussions are 1.029 and 1.068, respectively. In percentage terms, the abnormal trading volumes translate to 0.642 and 0.656, comparable to the figures in Brochet, Naranjo, and Yu [2016]. Consistent with Matsumoto, Pronk, and Roelofsen [2011], I find that $Volume^{Q\&A}$ is larger than $Volume^{PPT}$ (t-value of the mean difference = 4.04) and that $Return^{Q\&A}$ is larger than $Return^{PPT}$ (t-value of the mean difference = 12.83), indicating that there are more active market reactions during discussions than during presentations.

I report the variable of interest $Vocal\_Ambiguity$ separately at the presentation and discussion levels. The mean and standard deviation of $Vocal\_Ambiguity^{PPT}$ are 0.052 and 0.014, and the mean and standard deviation of $Vocal\_Ambiguity^{Q\&A}$ are 0.063 and 0.012, respectively. I test the statistical significance between the two average values and find that $Vocal\_Ambiguity^{Q\&A}$ is higher than $Vocal\_Ambiguity^{PPT}$ at the 1% level (t-value=104.51). This result indicates that the average vocal delivery is better during the presentation session than during the discussion session. This is consistent with Lee [2016] arguing that managers are likely to read pre-prepared scripts during presentation sessions but have to improvise their answers to some extent during discussion sessions.

Next, I report script-level variables *Fog*, *Grammar_Error*, *Audio_Length*, and *Tone* for presentation and discussion sessions separately. The average Fog index for the presentation session is 16.001, while the average Fog index for the discussion session is 11.876. The average audio length of the presentation session is 19.300 minutes, while the average length of the discussion session is 30.125 minutes. To facilitate the comparison, I report the number of grammatical errors scaled by the number of words for *Grammar_Error*.[17] The average number of grammatical errors scaled by the number of words during the presentation session is 0.005 and 0.013 for the discussion session. Also, the average net tonality of the presentation session is 0.007, while it is 0.003 for the discussion session. I conduct t-tests to validate the statistical significance of the differences in averages, and all differences are significant under 1% level. Combined, I find that presentation session speeches are more complex, last for a shorter period of time, contain fewer grammatical errors, and are more positive in terms of textual tone than discussion session speeches.

In Panel B, I report a correlation matrix for selected script-line level variables. I use script-line level observations during presentation sessions of earnings calls. As in Brochet, Naranjo, and Yu [2016], abnormal trading volume and abnormal absolute return are positively correlated with a Pearson correlation of 0.25. As expected, vocal ambiguity and abnormal absolute returns (trading volume) are negatively correlated, and the correlation is -0.04 (-0.06).

[17] I also calculate the number of abnormal article usages and the number of passive verbs. However, untabulated results indicate no statistical difference in these components between discussion and presentation sessions.

4 Empirical Results

4.1 Determinants of Vocal Ambiguity

To investigate the determinants of *Vocal_Ambiguity*, I test the following regression model:

$$Vocal\_Ambiguity_{qit} = \omega Z_{it} + \tau S_{qit} + \theta E_{qit} + FE + \epsilon_{qit} \tag{4}$$

I investigate the determinants of script-line level vocal ambiguity from presentation sessions of earnings calls. I include firm-level, script line-level, and executive-level control variables. As I require executive bio information from BoardEx, I have a reduced sample of 18,334.[18] I include year and industry fixed effects, and standard errors are clustered at the industry level.[19] To better illustrate the coefficients, I multiply 100 to $Vocal\_Ambiguity_{qit}$ in the regression.

Table 3 reports the regression results. As expected, individual characteristics are associated with vocal delivery. Vocal delivery improves when an executive is a native (-0.3759, t-value=-1.85) and is a male (-0.2270, t-value=-1.82).[20] However, vocal ambiguity is not associated with most firm-level variables, indicating that firm characteristics generally do not determine executives' vocal delivery. I show that vocal ambiguity is negatively associated with the number of analyst participants during the call (-0.0257, t-value=-2.21). This result im-

[18] Untabulated test using executive fixed effects instead of *Age*, *Native*, and *Gender* yields 104,951 observations. The regression results remain qualitatively similar.

[19] However, changing the clustering structure to speaker level does not affect my inferences.

[20] In terms of economic significance, male executives have an average of vocal ambiguity score which is 1.44 standard deviation (=0.2270/0.1573) lower than that of female executives. Also, native executives have an average vocal ambiguity score of 2.39 standard deviation (=0.3759/0.1573) lower than that of non-native executives. However, in my sample, only 8.3% of the executives are female, and 4.1% of the executives are non-native. Therefore, I acknowledge that the economic significance cannot be easily generalized.

plies that executives are likely to be clearer on their vocal delivery when there are many active listeners. This finding is consistent with the prior literature arguing that the vocal delivery of the same person may vary depending on the circumstances and training (Guiora, Brannon, and Dull [1972], Macdonald, Yule, and Powers [1994]). Furthermore, I find that vocal ambiguity is positively associated with delays in earnings calls (0.0076, t-value=2.33). Managers tend to withhold negative information by delaying earnings announcements (Begley and Fischer [1998]) and try to obfuscate investors with more complex disclosures (Bloomfield [2008]). Therefore, the executives may be making the speech sound more ambiguous to obfuscate investors during delayed earnings calls.

Next, regarding script-line level variables, I do not find a significant association between vocal ambiguity and Fog index (-0.0208, t-value=-1.33). In addition, there is no significant relation between vocal ambiguity and the frequency of grammatical errors (0.0459, t-value=1.38). These results jointly indicate that the vocal ambiguity measure could be independent of linguistic complexity and grammatical errors (Moustroufas and Digalakis [2007]). On the other hand, vocal ambiguity is negatively associated with the net tonality of speech (-11.3631, t-value=-5.27). When executives deliver a speech with positive content, they tend to be clearer in the vocal delivery. In terms of economic significance, a one standard deviation increase (0.0131) in net tonality leads to a 0.92 standard deviation decrease[21] in vocal ambiguity. Lastly, audio length is negatively associated with vocal ambiguity (-0.0004, t-value=-2.91),[22] implying that the

[21] The mean and standard deviation of script-line level tone are (0.008, 0.013). The mean and standard deviation of script line-level ambiguity multiplied by 100 are 5.3 and 0.16, respectively. Therefore, one standard deviation increase in tone leads to 11.3631×0.013/0.16 = 0.92 standard deviation decrease in vocal ambiguity.

[22] One-minute increase in audio length leads to 1.5 (=60×0.0004/0.016) standard deviation decrease in vocal ambiguity.

executives who are capable of maintaining longer speech also have better vocal delivery.

Taken together, individual characteristics such as gender and nationality primarily seem to affect vocal ambiguity score. When executives deliver a speech to a bigger audience, and when they deliver a positive-tone speech, their vocal delivery appears to improve. I do not find evidence that the measure of vocal ambiguity is associated with textual complexity and the frequency of grammar errors.

4.2 Univariate Analysis

Before I conduct regression analysis to test **H1**, I present univariate analysis results in Table 4. I first partition the presentation and discussion samples into quartiles depending on the value of *Vocal_Ambiguity*. Low *Vocal_Ambiguity* ($Vocal\_Ambiguity_{Q1}$) implies that the sample has superior vocal delivery, while high *Vocal_Ambiguity* ($Vocal\_Ambiguity_{Q4}$) implies that the sample has inferior vocal delivery. Then I present the mean and standard deviation of *Volume* and *Return* for $Vocal\_Ambiguity_{Q1}$ and $Vocal\_Ambiguity_{Q4}$ groups, respectively. Consistent with **H1**, I find that the mean abnormal trading volume and return are higher in the low vocal ambiguity group than in the high vocal ambiguity group. This result provides preliminary evidence of positive market reactions in response to better vocal delivery. The difference in $Volume^{PPT}$ between the two subgroups is 0.1106 (t-value = 5.79) and the difference in $Volume^{Q\&A}$ is 0.1632 (t-value = 8.68). The difference-in-differences is 0.0526 (p-value $<$ 0.0001), which indicates that the market reacts more sensitively to vocal ambiguity during discussions than during presentations. For $Return^{PPT}$ and $Return^{Q\&A}$,

I report the differences in average values of 0.0012 (t-value=6.35) and 0.0009 (t-value=4.66) between the two subgroups. However, the difference-in-differences is 0.0003 with a p-value of 0.1023 and is insignificant at a conventional level. This is probably because trading volume is the most visible market response, while the return is a joint product of investor reaction and trading position agreement among investors (Cready and Hurtt [2002]).

4.3 Effect of Vocal Ambiguity on Real-Time Market Reactions

In Table 5, I report the regression results to test **H1**. Specifically, I estimate Equation (1) by regressing *Vocal_Ambiguity* on *Volume* (*Return*) and controls. I include year and industry (SIC 2- digit classification) fixed effects, and standard errors are clustered at industry levels.[23] For all regression specifications, I find negative and statistically significant coefficients on *Vocal_Ambiguity*. These results indicate that vocal ambiguity is negatively associated with real-time abnormal market reactions, even after controlling for the various firm- and script text-related variables. In light of economic significance, a one standard deviation increase in vocal ambiguity score during presentation leads to a 2.36% (=0.014×1.7350/1.029) decrease in the average abnormal trading volume and a 2.69% (=0.014×0.0096/0.005) decrease in the average abnormal absolute returns. On the other hand, a one standard deviation increase in ambiguity score during discussion leads to a 3.66% (=0.012×3.2610/1.068) decrease in the average abnormal trading volume and a 3.42% (=0.012×0.0171/0.006) decrease in the average abnormal absolute returns. Comparing the magnitude of the decreases in market reactions, I observe that the market reacts more sensitively to

[23] Untabulated analyses using firm and industry level clusters yield qualitatively the same results.

vocal ambiguity during discussion sessions. To validate this observation, I compare the magnitude of the coefficients in Columns (1) and (3) and Columns (2) and (4), respectively. Indeed, the first F-test result reveals that β_1 of Column (3) is larger than that of Column (1) with a p-value less than 0.01. However, the second F-test results reveal that β_1 of Column (4) is not significantly larger than that of Column (2), with a p-value of 0.1722. This result is consistent with the preliminary observation in univariate analysis that vocal ambiguity affects abnormal trading volume more sensitively than it affects abnormal absolute returns.

Additionally, in Columns (3) and (4), I control for $Volume(Return)^{PPT}$ to minimize the effect of staggering market reactions. As expected, $Volume(Return)^{PPT}$ is positively associated with $Volume(Return)^{Q\&A}$. The adjusted R^2 of Column (3) is abnormally large (0.295), and this is because I include the lagged variable ($Volume^{PPT}$) as a control variable. In untabulated regression without $Volume^{PPT}$, the adjusted R^2 drops to 0.073, which is comparable to the R^2 of Column (1).

Other than vocal ambiguity, I find that the number of analysts following each firm and the number of earnings call participants are positively associated with real-time market reactions. Interestingly, *Grammar_Error* is negatively associated with abnormal volume (-0.0352, t-value = -3.81) and absolute returns (-0.0003, t-value = -2.65) only during the discussion sessions. This is because managers read written transcripts during presentation sessions and are less likely to commit grammar errors. Again, this finding is consistent with Brochet, Naranjo, and Yu [2016] which find that grammar errors are negatively associated with market reactions during discussion sessions.

In sum, the results provide strong support for **H1**, suggesting that when a call is acoustically ambiguous, market participants react negatively.

4.4 Script-Line Level Analysis on the Effect of Vocal Ambiguity on Market Reactions

Next, I turn to script-line level analysis. During each earnings call, generally more than one executive delivers speeches. Therefore, there is a within-call variation of vocal delivery. To examine whether the market immediately reacts to differing levels of vocal delivery, I test Equation (3) and present the results in Table 6. I use script lines during presentations delivered by executives and restrict the observations to be longer than 15 seconds. For $Volume(Return)^{PPT}$, I use real-time millisecond trade and quote data to capture the concurrent market reactions in response to each script line. This model specification provides us with a powerful setting to examine the direct effect of vocal delivery on investor behavior.

In Table 6, I first present script-line level regression results without control variables to illustrate the direct relation between *Vocal_Ambiguity* and $Volume(Return)^{PPT}$ (Column (1) and Column (3)). I include year and industry fixed effects, and standard errors are clustered at the industry level. I report the coefficient of -3.5827 (t-value = -6.35) in Column (1) and the coefficient of -0.0178 (t-value = -4.07) in Column (3). I find preliminary evidence from the results that vocal ambiguity is negatively associated with real-time investment decisions.

Next, in Column (2) and Column (4), I include call fixed effects in the regressions. With call fixed effects, I focus the analysis on the within-call varia-

tion of *Vocal_Ambiguity*.[24] Including call fixed effects subsumes other call-level characteristics that may affect real-time market reactions such as information contents from earnings announcements, information environment, or composition of investors. I exclude firm-level control variables and include script-line level controls as *Grammar_Error*, *Tone*, *Audio_Length*, and *Fog*. Also, I include executive fixed effects.

I also find strong evidence that *Vocal_Ambiguity* is negatively associated with $Volume(Return)^{PPT}$ even at the script line-level. Respectively, I report the coefficient of -6.8740 (t-value = -4.01) for Column (2) and the coefficient of -0.0477 (t-value = -3.31) for Column (4). In terms of economic significance, a one standard deviation increase in script-level vocal ambiguity leads to a 10.93% (=6.8740×0.0157/0.9868) decrease in average abnormal trading volume and a 26.74% (=0.0477×0.0157/0.0028) decrease in average absolute returns.[25] I find a stronger association between vocal ambiguity and market reaction variables in script-line level regressions than in call-level regressions. This is because the presentation or discussion-level vocal ambiguity score is a time-weighted average of the script line-level scores. Therefore, high and low ambiguity scores may cancel each other out during the calculation, leading to a lower econometric association.

The primary analysis includes script-line samples from presentation sessions. I acknowledge that script-line level observations from discussion sessions could be noisy since they are generally short and contain casual conversations between

[24] Call fixed effects account for 44.5% of the variation in vocal ambiguity, and within-call variation account for 55.5%.

[25] Per the suggestion of deHaan [2021], I also report economic significance based on a within-call variation of vocal ambiguity. One within-call standard deviation increase in script-level vocal ambiguity leads to a 5.71% (12.78%) decrease in average abnormal trading volume (absolute returns).

managers and analysts. As discussed earlier, script-line level *Vocal_Ambiguity* may yield biased values when assessing short speech fragments, and due to short audio length, investors may not have enough time to trade during each executive's speech. Nevertheless, untabulated regressions with call fixed effects show that *Vocal_Ambiguity* during discussion sessions is also negatively associated with abnormal trading volume (-1.8670, t-value = - 2.43). However, I do not see a statistically significant association between *Vocal_Ambiguity* and $Return^{Q\&A}$ under regressions with call fixed effects.

Overall, the results at the script line-level reinforce my inference that the market reacts *concurrently* to earnings call speech and reacts negatively in response to more acoustically ambiguous speeches.

4.5 Cross-Sectional Analyses

So far, I have shown that an ambiguous call results in a weaker market response. To better understand the impact of vocal delivery on market response, I examine the cross-sectional variation in the relation between vocal delivery and market reactions in this section.

4.5.1 Bad News Versus Good News

As discussed earlier, I expect the effect of vocal delivery on market reactions to be less pronounced when executives deliver negative news. This is because people input more cognitive resources when interpreting bad news and try to figure out the cause of such information (Soroka [2006]). Therefore, investors likely pay more attention to the contents of the calls and are less affected by ambiguous vocal delivery. I proxy for bad news with a negative earnings sur-

prise.[26]

Table 7, Panel A reports the results. Regarding the association between abnormal trading volume and vocal ambiguity, I find a statistically significant association when firms convey positive news (-1.9394, t-value = -2.54). In contrast, when firms convey negative news, the coefficient loses statistical significance (-0.7099, t-value = -0.74).[27] However, interestingly, I find a significant relation between vocal ambiguity and abnormal absolute returns under both subsamples (Column (2) and Column (4)). This is probably because investors react more uniformly in response to bad news. Even if there are not enough trades to generate a statistically significant relationship between the volume and vocal delivery, they may move in the same direction to yield a significant coefficient for Column (4). Taken together, the abnormal trading volume analysis supports **H2a**, while the abnormal absolute returns analysis provides weak evidence to support **H2a**.[28]

4.5.2 The Number of Analyst Participants During Discussion Session

I use the number of analyst participants during discussion sessions as a proxy for information demand of each earnings call (Hobson, Mayew, and Venkat-

[26]The second mechanism I propose, perception of investors, may also contribute to this conjecture. While the effect of positive perception toward the speaker delivering good news is clear, the effect for a speaker conveying bad news is not. If positive perception mitigates investors' negative reaction to bad news, the effect of vocal delivery on the market response will be smaller for negative news. However, I focus on the first mechanism, information processing costs, to focus on the theoretically grounded conjecture.

[27]When I conduct an F-test to compare the magnitudes of the two coefficients, I obtain a p-value of less than 0.01. This result indicates that the vocal delivery effect is more pronounced for firms with good news than those with bad news in terms of magnitude.

[28]To mitigate the concern that *Vocal_Ambiguity* score may differ significantly in the two groups, I perform a t-test to compare the mean values of *Vocal_Ambiguity* for both subsamples. I find that the mean values are 0.0519 and 0.0521 for the good news and bad news sample, respectively. These two values are not statistically different (t-value = -1.08).

achalam [2012], Brochet, Naranjo, and Yu [2016]).[29] I expect the vocal delivery effect to be more salient when there are many active listeners. This is simply because more active listeners use call information to make real-time investment decisions. I partition the sample using the median value of *N_Participant*, i.e., I classify observations with greater than or equal to six participants to be a high participation sample and vice versa.

Table 7, Panel B reports the results. As expected, I find a strong negative association between *Vocal_Ambiguity* and $Volume(Return)^{PPT}$ in high-participant subsample (Columns (1) and (2)). However, the coefficients lose their statistical significance in the low-participant subsample (Columns (3) and (4)).[30] Overall, the results support **H2b** that the demand-side of earnings call information also affects the magnitude of the effect.[31]

4.6 Sensitivity Checks

4.6.1 Alternative Measure of Vocal Delivery

To measure vocal delivery, I use *Vocal_Ambiguity* in the main analyses. However, to further validate the findings, I implement another vocal delivery measure conceptually similar to *Vocal_Ambiguity*, but calculated differently. Specifically, I calculate *Vocal_Clarity* by comparing the machine-retrieved texts

[29] However, unlike prior studies, I use the raw number of participants to partition the sample. This is because the intention is to capture the raw demand of information, not the scaled amount of information.

[30] I have 16,897 (16,409) samples in high-participant group and 12,900 (12,159) samples in the low-participant group. This imbalance occurs because *N_Participant* is discrete, and I set its median value as the partitioning point. Untabulated analysis using *N_Participant* = 5 as the partitioning point also yields similar results.

[31] I run F-tests to compare the coefficients' magnitudes. When I compare the coefficients of Columns (1) and (3) and the coefficients of Columns (2) and (4), I obtain p-values of less than 0.01, indicating that the coefficients of the high-participant group are larger than the coefficients in the low-participant group in terms of magnitude as well.

and edited transcripts provided by S&P Global.

First, I decompose the machine-retrieved texts into syllables and match them with pre-processed earnings call transcripts. I use an edited copy of S&P Global transcripts as pre-processed scripts. To create edited script copies, automated machine scripters record the calls, and then professional editors revise the outcomes. The editors do not add (delete) components to (from) machine-scripted versions of transcripts. But they edit punctuations or spellings that machines have difficulty translating (S&P [2017]). Thus, edited copies are generally free of punctuation or spelling errors but still contain grammatical errors.[32] Then, I compare and match the syllables of the retrieved text vectors and edited transcript vectors. The higher the similarity between the two vectors, the higher the delivery score is. I describe the mathematical details of the matching process in Online Appendix B2 (OA-B2).

Vocal_Clarity does not include any self-entropy calculation, but has a correlation coefficient of -0.63 (p-value < 0.01) with *Vocal_Ambiguity*. Hence, the two evaluation metrics conceptually yield similar results but are obtained differently. I substitute *Vocal_Ambiguity* in Equation (1) with *Vocal_Clarity* and replicate Table 5. Table 8, Panel A reports the results. Both in presentation and discussion sessions, I find a strong positive association between *Vocal_Clarity* and *Volume(Return)*. Furthermore, I find larger coefficients in discussion sessions (Column (3) and Column (4)), with F-test results that are statistically significant under a 0.01 confidence level. This finding is also consistent with the main analysis that the market reacts more sensitively during the discussion sessions than in presentation sessions. Additionally, untabulated tests show that

[32] Edited copies even include utterances. When presenters make a long pause or utterance (e.g., "um" sound in between the words), edited copies mark such sounds with "–".

script-line level regressions with call-level fixed effects yield similar results as in Table 6. Taken together, additional analyses using an alternative vocal delivery measure provide further support for the results of the main regressions.

4.6.2 Using One-Day Abnormal Trading Volume and Absolute Returns

Several prior studies that investigate the effect of earnings calls on market reactions use one- or two-day abnormal trading volume or abnormal absolute returns as dependent variables (Lee [2016]). I use real-time TAQ data in the main analyses to mitigate the concern arising from unknown omitted variables. However, I also implement one-day return and trading volume to validate the findings further. Dependent variable $Return_{it}^{0,1}$ refers to the market-adjusted abnormal absolute returns of firm i on the earnings call date of quarter t and $Volume_{it}^{0,1}$ refers to the market-adjusted abnormal trading volume of firm i on the earnings call date of quarter t. Independent variable $Vocal\_Ambiguity_{it}^{Call}$ indicates the weighted-average of *Vocal_Ambiguity* during the entire quarter t earnings call of firm i. Table 8, Panel B reports the results. I find no statistical evidence that $Vocal\_Ambiguity_{it}^{Call}$ is associated with $Volume(Return)_{it}^{0,1}$, implying that the vocal delivery effect lasts for a short period of time and gets automatically fixed throughout the day.

5 Conclusion

This study examines the relationship between vocal delivery during earnings calls and real-time market reactions. I employ a novel measure using a deep learning methodology to estimate the vocal delivery at the script line level. Furthermore, I obtain the timestamps for each line of earnings call transcripts,

which allows some rigorous tests on real-time market reactions in response to differing vocal delivery scores. Using a large sample of earnings calls from 2008 to 2020, I find strong evidence that high vocal ambiguity is associated with negative abnormal trading volume and abnormal absolute returns. This result holds even after controlling for various textual variables that affect investment decisions. In further analyses, I find that the vocal delivery effect is more pronounced when executives deliver good news and when there are more analyst participants in a call.

My study sheds light on the new dimension of how earnings call information affects investment decisions. Since earnings calls are audible, investors simultaneously integrate vocal and non-vocal information to reach their investment decisions. However, prior literature has been silent on the vocal dimension of the earnings call. I show that earnings call vocal delivery invokes real-time market reactions, even after controlling for textual information.

In current archival studies that examine the effect of information conveyance on investor behavior, differentiating linguistic complexity from content complexity has been an issue. In my setting, the vocal ambiguity measure is not economically associated with textual complexity or incorrect grammar usage. Therefore, by using the vocal delivery measure as the variable of interest, I clearly observe the effect of information conveyance on investment decisions. This study adds to this line of literature by uncovering that not only textual attributes but also audible contents affect investor behavior.

My research opens up a wealth of future research opportunities. In this study, even though I control for textual characteristics such as the Fog index or grammar errors, I do not measure the information contents of the executive

speech. It would be worthwhile to see to which content the market reacts more saliently in response to differing vocal delivery scores. I leave this open for future research.

References

J. Agnew and L. Szykman. Information overload and information presentation in financial decision making. In *Handbook of Behavioral Finance*. Edward Elgar Publishing, 2010.

H. S. Asay, R. Libby, and K. Rennekamp. Firm performance, reporting goals, and language choices in narrative disclosures. *Journal of Accounting and Economics*, 65(2-3):380–398, 2018.

A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

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

J. Begley and P. E. Fischer. Is there information in an earnings announcement delay? *Review of Accounting Studies*, 3(4):347–363, 1998.

J. Beshears, J. J. Choi, D. Laibson, B. C. Madrian, et al. *How does simplified disclosure affect individuals' mutual fund choices?* Number 13. University of Chicago Press Chicago, 2011.

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

E. Blankespoor, E. deHaan, and I. Marinovic. Disclosure processing costs, investors' information choice, and equity market outcomes: A review. *Journal of Accounting and Economics*, 70(2-3):101344, 2020.

R. Bloomfield. Discussion of "annual report readability, current earnings, and earnings persistence". *Journal of Accounting and Economics*, 45(2-3):248–252, 2008.

R. Bloomfield, F. Hodge, P. Hopkins, and K. Rennekamp. Does coordinated presentation help credit analysts identify firm characteristics? *Contemporary Accounting Research*, 32(2):507–527, 2015.

R. M. Bowen, A. K. Davis, and D. A. Matsumoto. Do conference calls affect analysts' forecasts? *The Accounting Review*, 77(2):285–316, 2002.

F. Brochet, P. Naranjo, and G. Yu. The capital market consequences of language barriers in the conference calls of non-us firms. *The Accounting Review*, 91

(4):1023–1049, 2016.

B. J. Bushee, I. D. Gow, and D. J. Taylor. Linguistic complexity in firm disclosures: Obfuscation or information? *Journal of Accounting Research*, 56(1):85–121, 2018.

A. C. Call, R. W. Flam, J. A. Lee, and N. Y. Sharp. Analysts' and managers' use of humor on public earnings conference calls. *Available at SSRN 3425509*, 2020.

S. Cao, W. Jiang, B. Yang, and A. L. Zhang. How to talk when a machine is listening: Corporate disclosure in the age of ai. Technical report, National Bureau of Economic Research, 2020.

Cision. Earnings call as a voluntary disclosure. 2017.

G. Cohen and A. Kudryavtsev. Investor rationality and financial decisions. *Journal of Behavioral Finance*, 13(1):11–16, 2012.

J. C. Cox, D. Kreisman, and S. Dynarski. Designed to fail: Effects of the default option and information complexity on student loan repayment. *Journal of Public Economics*, 192:104298, 2020.

W. M. Cready and D. N. Hurtt. Assessing investor response to information events using return and volume metrics. *The Accounting Review*, 77(4):891–909, 2002.

T. Davila and M. Guasch. Manager's body expansiveness, investor perceptions, and firm forecast errors and valuation. *Investor Perceptions, and Firm Forecast Errors and Valuation (April 27, 2021)*, 2021.

A. K. Davis, W. Ge, D. Matsumoto, and J. L. Zhang. The effect of manager-specific optimism on the tone of earnings conference calls. *Review of Accounting Studies*, 20(2):639–673, 2015.

E. deHaan. Using and interpreting fixed effects models. *Available at SSRN 3699777*, 2021.

R. Frankel, M. Johnson, and D. J. Skinner. An empirical examination of conference calls as a voluntary disclosure medium. *Journal of Accounting Research*, 37(1):133–150, 1999.

FRC. Video in corporate reporting. 2020.

A. Z. Guiora, R. C. Brannon, and C. Y. Dull. Empathy and second language learning 1. *Language Learning*, 22(1):111–130, 1972.

D. F. Gundersen and R. Hopper. Relationships between speech delivery and

speech effectiveness. *Communications Monographs*, 43(2):158–165, 1976.

J. L. Hobson, W. J. Mayew, and M. Venkatachalam. Analyzing speech to detect financial misreporting. *Journal of Accounting Research*, 50(2):349–392, 2012.

S. Hollander, M. Pronk, and E. Roelofsen. Does silence speak? an empirical analysis of disclosure choices during conference calls. *Journal of Accounting Research*, 48(3):531–563, 2010.

S. Hon-Snir, A. Kudryavtsev, and G. Cohen. Stock market investors: Who is more rational, and who relies on intuition. *International Journal of Economics and Finance*, 4(5):56–72, 2012.

L. Huang and H. Liu. Rational inattention and portfolio selection. *The Journal of Finance*, 62(4):1999–2040, 2007.

S. Huang, K. Snellman, and T. Vermaelen. Managerial trustworthiness and buybacks. *Journal of Financial and Quantitative Analysis*, pages 1–49, 2020.

M. Kacperczyk, S. Van Nieuwerburgh, and L. Veldkamp. A rational theory of mutual funds' attention allocation. *Econometrica*, 84(2):571–626, 2016.

M. D. Kimbrough. The effect of conference calls on analyst and market underreaction to earnings announcements. *The Accounting Review*, 80(1):189–219, 2005.

A. Lawrence, J. Ryans, E. Sun, and N. Laptev. Earnings announcement promotions: A yahoo finance field experiment. *Journal of Accounting and Economics*, 66(2-3):399–414, 2018.

J. Lee. Can investors detect managers' lack of spontaneity? adherence to predetermined scripts during earnings conference calls. *The Accounting Review*, 91(1):229–250, 2016.

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

D. Macdonald, G. Yule, and M. Powers. Attempts to improve english l2 pronunciation: The variable effects of different types of instruction. *Language Learning*, 44(1):75–100, 1994.

D. Matsumoto, M. Pronk, and E. Roelofsen. What makes conference calls useful? the information content of managers' presentations and analysts' discussion sessions. *The Accounting Review*, 86(4):1383–1414, 2011.

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

2012.

W. J. Mayew, M. Sethuraman, and M. Venkatachalam. Individual analysts' stock recommendations, earnings forecasts, and the informativeness of conference call question and answer sessions. *The Accounting Review*, 95(6): 311–337, 2020.

J. Mercer. Prospect theory and political science. *Annual Review of Political Science*, 8:1–21, 2005.

N. Moustroufas and V. Digalakis. Automatic pronunciation evaluation of foreign speakers using unknown text. *Computer Speech & Language*, 21(1): 219–230, 2007.

A. Niewiadomski and A. Akinwale. Efficient similarity measures for texts matching. 2015.

S. M. Price, J. S. Doran, D. R. Peterson, and B. A. Bliss. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, 36(4):992–1011, 2012.

K. Rennekamp. Processing fluency and investors' reactions to disclosure readability. *Journal of Accounting Research*, 50(5):1319–1354, 2012.

A. Saporta, T.-H. Vu, M. Cord, and P. Pérez. Esl: Entropy-guided self-supervised learning for domain adaptation in semantic segmentation. *arXiv preprint arXiv:2006.08658*, 2020.

A. K. Shah and D. M. Oppenheimer. Easy does it: The role of fluency in cue weighting. 2007.

H. A. Simon. Bounded rationality. In *Utility and Probability*, pages 15–18. Springer, 1990.

S. N. Soroka. Good news and bad news: Asymmetric responses to economic information. *The Journal of Politics*, 68(2):372–385, 2006.

S&P. Earnings conference call and transcript. 2017.

R. Srikanth, B. Li, and J. Salsman. Automatic pronunciation scoring and mispronunciation detection using cmusphinx. In *Proceedings of the Workshop on Speech and Language Processing Tools in Education*, pages 61–68, 2012.

S. C. Tasker. Bridging the information gap: Quarterly conference calls as a medium for voluntary disclosure. *Review of Accounting Studies*, 3(1):137–167, 1998.

K. J. Van Engen and J. E. Peelle. Listening effort and accented speech. *Frontiers*

in Human Neuroscience, 8:577, 2014.

K. J. Van Engen, B. Chandrasekaran, and R. Smiljanic. Effects of speech clarity on recognition memory for spoken sentences. 2012.

A. Vrij. *Detecting lies and deceit: Pitfalls and opportunities*. John Wiley & Sons, 2008.

C. M. Ward, C. S. Rogers, K. J. Van Engen, and J. E. Peelle. Effects of age, acoustic challenge, and verbal working memory on recall of narrative speech. *Experimental Aging Research*, 42(1):97–111, 2016.

B. Weiner. An attributional theory of achievement motivation and emotion. *Psychological review*, 92(4):548, 1985.

S. M. Witt and S. J. Young. Language learning based on non-native speech recognition. In *Eurospeech*. Citeseer, 1997.

S. Wu, C. Tan, J. Kleinberg, and M. Macy. Does bad news go away faster? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 646–649, 2011.

G. L. Zahn. Cognitive integration of verbal and vocal information in spoken sentences. *Journal of Experimental Social Psychology*, 9(4):320–334, 1973.

Y. Zou, Z. Yu, B. Kumar, and J. Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 289–305, 2018.

M. Zuckerman, B. M. DePaulo, and R. Rosenthal. Verbal and nonverbal communication of deception. In *Advances in Experimental Social Psychology*, volume 14, pages 1–59. Elsevier, 1981.

Online Appendix A (OA-A). Retrieving Texts and Obtaining Timestamps

This section describes how I obtain timestamps for each script line. I first infer audio files into texts by performing the following steps.

1. I download the audio files in mp3 format from S&P Global.

2. By processing each file with `Pydub` package in Python, I convert the files into *wav* format. Also, I set the sampling rate of each audio file to be 16,000 frames per second.

3. Soundfile package in Python converts audio files into number arrays.

4. `Wav2Vec2.0` (Baevski et al. [2020]) is a pre-trained sound inference model that converts sound waves into texts. It achieves a state-of-the-art accuracy rate in classifying sound wave vector data into corresponding syllables. I use `Wav2Vec2.0` for CTC (Connectionist Temporal Classification)[33] from Huggingface Transformer to convert the wave data into syllables.

5. I assign 50,000 frames to each input group. Since I have 16,000 frames per second, each input group is approximately 3.125 seconds.[34]

6. `Wav2Vec2.0ForCTC` assigns a "logit value vector" that expresses the relative probability that an input group is assigned to each syllable. It classifies the number array into one of the 32 distinct syllables. For instance, if the wave function is clear to be translated into the syllable "ha",

[33] CTC is a type of neural network-based scoring function that is used to classify the data into several categories, especially when the timing is variable. For instance, CTC is used for data that changes continuously over time, such as classifying handwritten texts by a cell phone user on the screen. In my task, since audio file wave data is constantly changing over the length of each audio, I use a CTC classifier to identify the corresponding syllable for each time frame.

[34] If I assign too many frames per second, it becomes more likely that a single word is divided between two input groups. For instance, if a speaker pronounces "integrity" from 3.10 seconds to 4.05 seconds, the word will be assigned to two different input groups, being a potential source of execution error. In contrast, if I assign less frames per second into a single group, the accuracy of timestamping deteriorates. I experiment with various frames per second and find 50,000 frames as one of the optimal options. Therefore, the maximum error rate for the timestamp that I obtain is 3.125 seconds.

`Wav2Vec2.0ForCTC` assigns a high logit value to the syllable "ha" and low values to the other 31 syllables. However, if the wave is somehow contaminated, the classifier cannot assign a single syllable to the wave vector. Instead, it sets a logit vector assigning relative probabilities to the wave vector. If the vector resembles the wave vectors of syllables "ha", "ah", and "la" at the same time, the classifier will assign high probabilities to those three syllables and near-zero to the other 29 syllables.

7. By decoding the logit vector of each input group, I retrieve the texts from audio files.

Now, I have input groups (3.125 seconds each) and their corresponding retrieved texts. In the script data set, each line denotes a single speaker. When there is a change in speaker, the script moves to the next line. I aim to obtain timestamp (start time and end time) for each script line.

Since the retrieved texts are not perfect, the texts cannot be matched directly to the edited copies. Furthermore, edited copies intentionally omit some of the speech fragments as short conversations or operator instructions. Therefore, I use n-gram to match the retrieved texts with edited copies provided by S&P Global. n-gram is a representative method to match texts based on their textual similarity and a number of natural language processors utilize this method (Niewiadomski and Akinwale [2015]). In my research, I use character-level 4-gram matching. If I decompose "Good morning" into character-level 4-gram components, I obtain "Good", "ood ", "od m", "d mo", " mor", "morn", "orni", "rnin", and "ing" (space is also a gram). I take the following steps to obtain the timestamp for each script line.

8. I set a window with the same length with each input group. I slide the window in the actual transcript and obtain 4-gram components. I then match the 4-gram components of the input group with the components of rolling windows. I obtain the number of matched 4-gram components and scale it with the number of 4-gram components in the input group (4-gram score). Therefore, I obtain one 4-gram score that corresponds to one window location.

9. I calculate the 4-gram score by sliding the window in the current and

following script lines.

10. I then set a "pointer" that moves according to the following set of rules:

    i. Based on the max(4-gram score) of the two script lines I investigate in 9, if the max(4-gram score) is higher in the previous line, the pointer stays. However, if the max(4-gram score) is higher in the following line, the pointer moves to the next line.

    ii. I calculate the text length of each script line (A) and the text length of the input groups that are already matched to that script line (B). Let the length of the current input group be C. If $(B + C) < 0.6A$, I force the pointer to stay. However, if $(B + C) > 1.2A$, I force the pointer to move to the following line.

    iii. However, one trivial problem is that the edited scripts omit operator instructions. Sometimes, these instructions are sufficiently long to evade the exception rule (ii). Therefore, I relax the exception rule and apply the time-based rule to the pointer. Now, I calculate the audio length in seconds of the input groups that are already matched to that script line (B^t). If there is an indicator "[Operator Instructions]" or "[Operator]" in the script line, I force the pointer to stay if $B^t < 10$. On the other hand, I force the pointer to move if $B^t > 300$.

11. Since each input group is 3.125-second long, I can calculate the start-time and end-time of each script-line.

Online Appendix B1 (OA-B1). Calculating *Vocal_Ambiguity*

In evaluating vocal delivery, the most commonly used measure is the Goodness of Pronunciation (GoP) (Witt and Young [1997]). However, this measure combines (i) phoneme delivery clarity and (ii) its similarity to the predetermined "answers." I do not use this evaluation scheme in my research for the following reasons. First, the scoring system requires a "perfect phoneme sequence," or a perfect transcript of the audio files. However, the edited scripts omit several audio components such as instructions or short casual talks. Second, GoP assigns high scores to the pronunciation that assembles the pronunciation of the natives (answers). However, I aim at assessing the vocal delivery of the speakers, not at determining whether the executive has a native accent or not. Therefore, this study uses the first component (phoneme delivery clarity) as the primary vocal delivery measure.

Vocal_Ambiguity is a self-entropy-based phoneme delivery measure that can be calculated from the logit vector of OA-A, step 6. Self-entropy is a measure of probabilistic confidence (Zou et al. [2018], Saporta et al. [2020]). In classifying tasks, a deep-learning-based classifier assigns a probability vector to each input. The vector represents a relative probability that the input will be assigned to each category. For instance, if the classifier performs a task to assign each observation into one of 10 categories, it calculates a vector $(p_1, p_2, \cdots, p_{10})$ for each observation. p_i refers to the probability that the observation belongs to category i. If the observation is clearly type i, the classifier will assign a high probability to p_i and assign low probabilities to p_j $(i \neq j)$. In contrast, if the observation is rather confusing, the classifier will assign even probabilities to several p_i's.

The generalized calculation of self-entropy is as follows (OA-(1)):

$$Entropy(s_i) = -\frac{1}{\log M} \sum_{k=0}^{M-1} l_k(s_i) \log l_k(s_i) \tag{5}$$

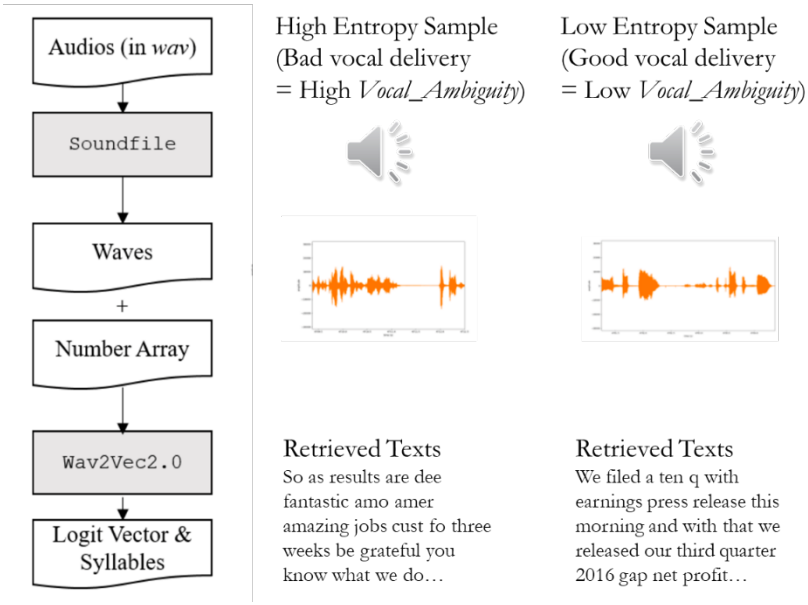
,where s_i denotes an input, M denotes the number of categories, and $l_k(s_i)$ denotes the logit probability that s_i is assigned to category k. In the research design, $M = 32$. Low self-entropy indicates that an input is assigned to a certain category with mathematical certainty. On the other hand, high self-entropy

implies that the classifier is unable to assign the input into a sole category. Therefore, low entropy in syllable classification task implies that the vocal delivery is clear. On the other hand, high entropy means that the wave vector is confusing or the vocal delivery is ambiguous. I use the average entropy of each script-line as $Vocal\_Ambiguity_{qit}$. Refer to Figure OA-1, for visual illustration.

In Figure OA-1, I briefly pipeline the algorithm used to retrieve the texts. I provide two sample input groups with high entropy (0.0623) and low entropy (0.0081). I visually present the converted waves and number arrays. As expected, the high entropy sample shows poor text retrieval, while the low entropy sample shows near-perfect text retrieval.

Figure OA-1. Text Retrievals with High and Low Entropy Input Group

This figure illustrates the pipeline of text retrieval algorithms. The column on the left-hand side points out the algorithm flow along with the program's Python packages. The column on the right-hand side provides two actual examples in the earnings call audios. Each sample is a 3.125-second input group with differing entropy scores (0.0788 and 0.0101, respectively). The last row shows the converted texts from the input groups.



Online Appendix B2 (OA-B2). Calculating *Vocal_Clarity*

I calculate *Vocal_Clarity* from OA-A, step 10 with the following two steps.

1. I calculate the number of matched 4-grams for each 3.125-second input group. I obtain the match score (*Match_Score*) by scaling the number of matched 4-grams with the total number of 4-grams in each input group.

2. After the timestamping steps, I have each input group assigned to a script line. A script line is likely to have more than two matched input groups. I take the averaged value of *Match_Score* of the matched input groups to calculate a script-line level 4-gram match score (*Vocal_Clarity*).

Refer to Table OA-1 for the visual representation of well-retrieved and poorly retrieved texts.

Table OA-1. Visual Representation of Text Retrievals Depending on 4-Gram Scores

This table reports two examples of text retrievals based on their 4-gram score. The upper row is an example of perfect retrieval (4-gram = 1.000), and the second row is an example of poor retrieval (4-gram = 0.371). Even though the second line fails to retrieve most of the words, they "sound" similarly to the edited script. This result confirms that the scripting algorithm functions well to capture the core "sound" even though the speaker has poor vocal delivery.

4-gram	Edited Script (One sentence)	Retrieved Text (One sentence)
1.000	Mostly related to the lower purchasing price of corn.	Mostly related to the lower purchasing price of corn.
0.371	They want to stay local, they don't need the money.	The on o tey loc they don't need e mon.

Appendix A. Variable descriptions

Variable Name	Description	Source
Dependent variables		
$Volume^{PPT(Q\&A)}$	Abnormal trading volume during the presentation (discussion) section of earnings call. Refer to Section 3 for detailed calculation.	TAQ
$Return^{PPT(Q\&A)}$	Abnormal absolute returns during the presentation (discussion) section of earnings call. Refer to Section 3 for detailed calculation.	TAQ
Independent variables		
Vocal_Ambiguity	The degree of earnings call delivery measured by *self-entropy*. Refer to Online Appendix B1 for a detailed variable definition.	S&P Global
Firm-level variables		
Size	The natural logarithm of the market value of equity.	Compustat
ROA	Net income divided by total assets.	Compustat
MTB	Market value of equity divided by total assets.	Compustat
Leverage	Total debt divided by total assets.	Compustat
Earnings_Volatility	Standard deviation of the net income over the past five quarters.	Compustat
Loss	Indicator that equals one if a firm reports negative net income for the quarter and zero otherwise.	Compustat
N_Seg	The number of business segment at the end of fiscal year.	Compustat segment
M&A	Indicator that equals one if a firm has announced mergers and acquisitions during the quarter and zero otherwise.	SDC Platinum
N_Analyst	The number of analysts following a firm at the end of each quarter.	I/B/E/S
SUE	Absolute difference of between the median analyst forecast EPS and actual EPS, scaled by actual EPS.	I/B/E/S

Return_Volatility	Return volatility of the daily returns of the past quarter.	CRSP
Return_Sign	Indicator that equals one if the return of the following quarter is positive and zero otherwise.	CRSP
Friday	Indicator that equals one if a call is held on Friday and zero otherwise.	S&P Global
N_Participants	The number of analyst participants during a call.	S&P Global
Lag	The period between the fiscal quarter end and earnings call date in days.	S&P Global
Fourth_Q	Indicator that equals one if a call is held in the fourth fiscal quarter of a firm and zero otherwise.	S&P Global
OneDayRet	Abnormal return one day before the call.	CRSP
Script-level variables		
Fog	Fog index of the script.	S&P Global
Audio_Length	The length of the earnings call audio in seconds.	S&P Global
Grammar_Error	The average of the following three items following Brochet, Naranjo, and Yu [2016]: (i) standardized score of the abnormal use of article 'the', (ii) standardized score of the abnormal use of passive voice, (iii) standardized score of grammatical errors other than punctuation.	S&P Global
Tone	Net tone of the script calculated by subtracting the number of negative words from the number of positive words, scaled by the number of total words. I use the financial vocabulary dictionary provided by Loughran and McDonald [2011].	S&P Global
Executive-level variables		
Age	Age of the speaker in years.	BoardEx
Gender	Indicator that equals one if a speaker is male and zero otherwise.	BoardEx

Native	Indicator that equals one if a speaker is from America, Canada, Australia, and England, and zero otherwise.	BoardEx

Table 1. Sample Reconciliation

This table reports the sample reconciliation procedures. #Calls denotes the number of distinct calls and #Lines denotes the number of script lines that I use for analysis. Each line is assigned to one speaker, i.e. when the speaker changes, the script also moves to the next line. I download all available earnings call transcripts from S&P Global Capital IQ Transcript database. Since I aim at investigating real-time market reactions during earnings calls, I only count the calls that happen during trading hours. Then I match each script with its corresponding audio file. I lose several scripts due to audio file inavailability such as truncated audios, missing audios, and extremely short audios. Furthermore, I require control variables from Compustat, CRSP, and I/B/E/S databases.

	# Calls	# Lines
Earnings call scripts of U.S. firms available from S&P Global Capital IQ (2008 – 2020)	119,211	7,031,038
Less:		
Calls without valid audio files	(1,165)	(81,526)
Calls that happen when the stock market is closed	(86,633)	(4,941,953)
Short calls (calls less than 21.78 minutes)	(5,644)	(174,774)
Lines during discussion sessions	-	(1,814,023)
Lines from operator and lines that are less than 15 seconds	-	(65,915)
Insufficient financial data	(116)	(304)
Total sample	30,224	107,316

Table 2. Descriptive Statistics

This table reports the descriptive statistics of the variables used in regressions. Panel A reports the number of observations, mean, standard deviation, Q1 and Q3 for each variable. For *Grammar_Error*, I use the value before the standardization to facilitate the comparison between presentation and discussion sessions. Panel B reports correlation matrices. Each panel reports the Pearson (below the diagonal) and Spearman (above the diagonal) correlation among script line-level variables. Each variable is measured at script-line level during presentation sessions. Here, change in script lines indicates a change in speakers during the calls. Bold numbers indicate statistical significance at 10% level.

Panel A. Descriptive Statistics						
	# Obs.	Mean	Std.	Q1	Median	Q3
Market Reactions						
$Volume^{PPT}$	29797	1.029	1.174	0.336	1.028	1.740
$Return^{PPT}$	28568	0.005	0.011	-0.001	0.002	0.007
$Volume^{Q\&A}$	29774	1.068	1.133	0.404	1.072	1.752
$Return^{Q\&A}$	28525	0.006	0.012	-0.001	0.003	0.009
Delivery						
$Vocal\_Ambiguity^{PPT}$	30224	0.052	0.014	0.042	0.049	0.058
$Vocal\_Ambiguity^{Q\&A}$	30126	0.063	0.012	0.055	0.061	0.069
Firm Variables						
Size	30224	7.415	1.771	6.235	7.450	8.575
ROA	30224	0.005	0.034	0.001	0.008	0.019
MTB	30224	2.055	0.253	1.921	2.088	2.220
Leverage	30224	0.619	0.244	0.452	0.616	0.802
Earnings_Volatility	30224	62.974	158.776	3.880	12.019	41.860
Loss	30224	0.214	0.410	0.000	0.000	0.000
N_Seg	30224	5.672	5.319	1.000	3.000	9.000
M&A	30224	0.015	0.123	0.000	0.000	0.000
N_Analyst	30224	9.026	6.820	4.000	7.000	13.000
SUE	30224	0.298	0.688	0.022	0.081	0.250
Return_Volatility	30224	0.022	0.016	0.013	0.018	0.027
Return_Sign	30224	0.527	0.499	0.000	1.000	1.000
Script Variables						
Friday	30224	0.143	0.350	0.000	0.000	0.000
N_Participants	30224	6.691	3.706	4.000	6.000	9.000
Lag	30224	34.063	11.942	26.000	32.000	39.000
OneDayRet	30224	0.000	0.077	-0.039	0.000	0.039
Fourth_Q	30224	0.244	0.429	0.000	0.000	0.000
Script_Presentation						
Fog	30224	16.001	1.556	14.992	16.000	17.024
Grammar_Error	30224	0.005	0.003	0.002	0.004	0.006
Audio_Length	30224	19.300	6.756	14.427	18.594	23.438
Tone	30224	0.007	0.007	0.003	0.007	0.012
Script_Discussion						
Fog	30126	11.876	1.058	11.121	11.806	12.551
Grammar_Error	30126	0.013	0.004	0.011	0.013	0.015
Audio_Length	30126	30.125	12.347	20.573	29.688	38.542
Tone	30126	0.003	0.005	-0.001	0.003	0.006

Executive						
Age	18334	57.106	8.459	52.000	57.000	62.000
Native	18334	0.959	0.197	1.000	1.000	1.000
Gender	18334	0.917	0.276	1.000	1.000	1.000

Panel B. Correlation Matrix (Script Line-level Variables)

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
[1] *Volume*		**0.17**	**-0.04**	**0.01**	**-0.05**	**0.04**	**0.02**	-0.01	**0.02**	**-0.03**
[2] *Return*	**0.25**		**-0.03**	**0.03**	**-0.05**	**0.10**	**-0.02**	0.00	0.01	0.01
[3] *Vocal_Ambiguity*	**-0.04**	**-0.06**		**-0.04**	**0.12**	**-0.16**	**-0.10**	**-0.04**	**-0.04**	**-0.05**
[4] *Fog*	**0.01**	**0.03**	**-0.05**		**-0.10**	**0.13**	**0.04**	**-0.03**	**-0.07**	0.00
[5] *Grammar_Error*	**-0.04**	**-0.05**	**0.13**	**-0.08**		-0.45	**-0.06**	**0.05**	**0.03**	0.00
[6] *Audio_Length*	**0.04**	**0.11**	**-0.19**	**0.19**	**-0.48**		**-0.05**	0.00	-0.01	0.01
[7] *Tone*	**0.03**	**-0.01**	**-0.13**	**0.03**	**-0.06**	**0.07**		**0.03**	**0.02**	**-0.06**
[8] *Age*	**-0.02**	-0.01	**-0.05**	**-0.02**	**0.04**	-0.01	0.05		**0.12**	**0.07**
[9] *Gender*	**0.02**	0.01	**-0.05**	**-0.07**	**0.04**	0.00	**0.03**	**0.13**		**-0.05**
[10] *Native*	**-0.03**	0.00	**-0.05**	-0.01	0.00	**0.02**	**-0.05**	**0.08**	**-0.05**	

Table 3. Determinants of Earnings Call Delivery

*, **, *** represent statistical significance at the 10 percent, 5 percent, and 1 percent levels, respectively.

This table reports the OLS regression results on the relation between pronunciation of each executive and other variables. $Vocal\_Ambiguity_{qit}$ denotes the delivery score of executive q's speech during the presentation of firm i's earnings call in year t. To better represent the magnitude of regression coefficient, I multiply 100 to raw $Vocal\_Ambiguity_{qit}$. Refer to Appendix A for detailed variable descriptions. Standard errors are clustered by industry. All continuous variables are winsorized at the 1 percent and 99 percent levels. t-values are reported in the parenthesis.

	$Vocal\_Ambiguity_{qit}$
Firm-level variables	
$Size_{it}$	-0.0133
	(-0.16)
ROA_{it}	-0.5662
	(-0.49)
MTB_{it}	-0.1325
	(-0.28)
$Leverage_{it}$	0.0725
	(0.37)
$Earnings\_Volatility_{it}$	-0.0002
	(-0.73)
$Loss_{it}$	-0.1037
	(-1.50)
N_Seg_{it}	0.0047
	(0.53)
$M\&A_{it}$	0.0499
	(0.43)
$N\_Analyst_{it}$	0.0021
	(0.22)
SUE_{it}	0.0086
	(0.33)
$Return\_Volatility_{it}$	-0.3641
	(-0.27)
$Return\_Sign_{it}$	0.0365
	(1.33)
$Friday_{qit}$	-0.0916
	(-1.45)
$N\_Participant_{qit}$	-0.0257**
	(-2.21)
Lag_{qit}	0.0076**
	(2.33)
$Fourth\_Q_{qit}$	-0.0006
	(-0.01)
$OneDayRet_{qit}$	-0.1617
	(-0.83)
Script line-level variables	
$Grammar\_Error_{qit}$	0.0522

	(1.50)
$Audio\_Length_{qit}$	-0.0004***
	(-2.91)
Fog_{qit}	-0.0208
	(-1.33)
$Tone_{qit}$	-11.3631***
	(-5.27)
Executive-level variables	
$Native_{qit}$	-0.3759*
	(-1.85)
Age_{qit}	0.0012
	(0.20)
$Gender_{qit}$	-0.2270*
	(-1.82)
Year-Fixed Effect	Yes
Industry-Fixed Effect	Yes
Adjusted R^2	0.102
# Obs.	18,334

Table 4. Univariate Analysis

*, **, *** represent statistical significance at the 10 percent, 5 percent, and 1 percent levels, respectively.

This table reports the univariate analysis results. I partition the delivery score of presentation and discussion sections of each earnings call into quartiles. Then I examine the difference in the mean values of abnormal trading volume and the absolute value of abnormal return during the presentations and discussions, depending on the delivery score. All continuous variables are winsorized at the 1 percent and 99 percent levels. t-values are reported in the parenthesis.

	$Vocal\_Ambiguity_{Q1}$		$Vocal\_Ambiguity_{Q4}$		Difference
	Mean	Std	Mean	Std	
	(1)	(2)	(3)	(4)	(1) – (3)
$Volume^{PPT}$	1.0988	1.0746	0.9882	1.2505	0.1106***
					(5.79)
$Return^{PPT}$	0.0053	0.0111	0.0041	0.0104	0.0012***
					(6.35)
$Volume^{Q\&A}$	1.1643	1.1054	1.0012	1.1853	0.1632***
					(8.68)
$Return^{Q\&A}$	0.0064	0.0122	0.0055	0.0117	0.0009***
					(4.66)

Table 5. Effect of Earnings Call Vocal Ambiguity on Real-Time Market Reactions

*, **, *** represent statistical significance at the 10 percent, 5 percent, and 1 percent levels, respectively.

This table reports the regression results of the relation between earnings call vocal delivery and market reactions. Superscript PPT and $Q\&A$ denote presentation and discussion session, respectively. Fog_{it}, $Audio\_Length_{it}$, and $Tone_{it}$ are session-level variables (measured separately for each presentation and discussion). Refer to Appendix A for detailed variable descriptions. Columns (1) and (2) report the real-time market reactions during the presentation session of the calls and Columns (3) and (4) report the real-time market reactions during the discussion session of the calls. Standard errors are clustered by industry. All continuous variables are winsorized at the 1 and 99 percent levels. t-values are reported in the parenthesis.

	(1) $Volume_{it}^{PPT}$	(2) $Return_{it}^{PPT}$	(3) $Volume_{it}^{Q\&A}$	(4) $Return_{it}^{Q\&A}$
$Vocal\_Ambiguity_{it}$	-1.7350***	-0.0096*	-3.2610***	-0.0171**
	(-3.22)	(-1.78)	(-5.67)	(-2.27)
Firm-level controls				
$Size_{it}$	0.0305	-0.0010***	-0.0349	-0.0009***
	(0.69)	(-3.36)	(-1.05)	(-2.91)
ROA_{it}	-0.2020	-0.0076*	0.6290**	0.0078*
	(-0.71)	(-1.91)	(2.03)	(1.95)
MTB_{it}	0.0807	0.0046***	0.3350	0.0049***
	(0.32)	(3.91)	(1.65)	(3.04)
$Leverage_{it}$	-0.0483	0.0007	-0.0241	0.0001
	(-0.42)	(0.74)	(-0.35)	(0.13)
$Earnings\_Volatility_{it}$	0.0001	0.0000	0.0000	0.0000
	(1.15)	(1.15)	(0.76)	(0.54)
$Loss_{it}$	-0.0813**	-0.0003	-0.0164	0.0001
	(-2.56)	(-0.98)	(-0.69)	(0.27)
N_Seg_{it}	0.0017	0.0000	0.0014	0.0000
	(0.52)	(0.34)	(0.38)	(0.36)
$M\&A_{it}$	-0.0511	0.0000	0.0942**	0.0008
	(-1.38)	(0.02)	(2.21)	(0.97)
$N\_Analyst_{it}$	-0.0003	-0.0000	-0.0029*	-0.0000
	(-0.11)	(-0.43)	(-1.77)	(-0.03)
SUE_{it}	0.0714***	0.0002*	0.0382***	0.0002
	(4.73)	(1.91)	(3.79)	(1.00)
$Return\_Volatility_{it}$	-4.4720***	0.0053	-1.9440***	0.0162**
	(-6.69)	(0.81)	(-3.26)	(2.05)
$Return\_Sign_{it}$	-0.0414**	0.0001	-0.0159	-0.0002
	(-2.35)	(1.05)	(-0.98)	(-1.08)
$Friday_{it}$	-0.0059	0.0002	0.0320	-0.0001
	(-0.25)	(0.81)	(1.24)	(-0.70)
$N\_Participant_{it}$	0.0225***	0.0000546	0.00595*	0.0000242
	(6.89)	(1.41)	(1.95)	(0.53)
Lag_{it}	0.0032***	0.0000	0.0011	0.0000***
	(2.88)	(0.36)	(1.43)	(2.71)

$Fourth\_Q_{it}$	-0.0026	-0.0000	0.0045	-0.0003
	(-0.13)	(-0.44)	(0.30)	(-1.32)
$OneDayRet,_{it}$	-0.0903	-0.0056***	-0.1010	-0.0043**
	(-0.71)	(-5.41)	(-1.31)	(-2.57)
$Volume\ (Return)^{PPT}{}_{it}$			0.4400***	0.1690***
			(28.14)	(10.87)
Script-level controls				
Fog_{it}	-0.0093	-0.0000	-0.0027	-0.0001
	(-1.04)	(-0.21)	(-0.36)	(-0.64)
$Grammar\_Error_{it}$	-0.0355	-0.0007*	-0.0058	-0.0001
	(-0.86)	(-1.96)	(-0.51)	(-0.93)
$Audio\_Length_{it}$	0.0088***	0.0001***	0.0017**	0.0001***
	(4.59)	(7.39)	(2.15)	(7.33)
$Tone_{it}$	-3.0630**	-0.0273**	-1.0660	-0.0246
	(-2.37)	(-2.49)	(-0.73)	(-1.19)
Year-Fixed Effect	Yes	Yes	Yes	Yes
Industry-Fixed Effect	Yes	Yes	Yes	Yes
Adjusted R^2	0.147	0.096	0.328	0.110
# Obs.	29,797	28,658	29,774	28,525

Table 6. Speaker-level Regressions with Call Fixed Effects During Presentations

*, **, *** represent statistical significance at the 10 percent, 5 percent, and 1 percent levels, respectively.

This table reports the speaker-level OLS regression results. I obtain speaker-level delivery score during the presentation session of each earnings call and the corresponding market reactions when each executive is speaking. $Vocal\_Ambiguity_{qit}$ denotes the delivery score of script line q of firm i in year t earnings call. When including call-fixed effects, I include the variables that can be calculated at a script-line level as *Grammar_Error*, *Tone*, *Audio_Length*, and *Fog*. Standard errors are clustered by industry. All continuous variables are winsorized at the 1 percent and 99 percent levels. t-values are reported in the parenthesis.

	(1)	(2)	(3)	(4)
	$Volume_{qit}^{PPT}$	$Volume_{qit}^{PPT}$	$Return_{qit}^{PPT}$	$Return_{qit}^{PPT}$
$Vocal\_Ambiguity_{qit}$	-3.5827***	-6.8740***	-0.0178***	-0.0477***
	(-6.35)	(-4.01)	(-4.07)	(-3.31)
Firm-Level Controls	No	No	No	No
Script Line-Level Controls	No	Yes	No	Yes
Executive-Fixed Effect	No	Yes	No	Yes
Year-Fixed Effect	Yes	No	Yes	No
Industry-Fixed Effect	Yes	No	Yes	No
Call-Fixed Effect	No	Yes	No	Yes
Adjusted R^2	0.038	0.373	0.016	0.195
# Obs.	94,569	89,229	78,151	71,359

Table 7. Cross-Sectional Tests

*, **, *** represent statistical significance at the 10 percent, 5 percent, and 1 percent levels, respectively.
Panel A reports the firm-level effect of bad (good) news on the relation between earnings call delivery and immediate market reactions. Panel B reports the effect of the number of earnings call participants on the relation between earnings call delivery and real-time market reactions. Standard errors are clustered by industry. All continuous variables are winsorized at the 1 percent and 99 percent levels. t-values are reported in the parenthesis.

Panel A. Good News Versus Bad News				
	Good News		Bad News	
	(1)	(2)	(3)	(4)
	$Volume_{it}^{PPT}$	$Return_{it}^{PPT}$	$Volume_{it}^{PPT}$	$Return_{it}^{PPT}$
$Vocal\_Ambiguity_{it}$	-1.9394***	-0.0119**	-0.7099	-0.0180**
	(-2.54)	(-2.13)	(-0.74)	(-3.28)
Firm-Level Control	Yes	Yes	Yes	Yes
Script-Level Control	Yes	Yes	Yes	Yes
Year-Fixed Effect	Yes	Yes	Yes	Yes
Industry-Fixed Effect	Yes	Yes	Yes	Yes
Adjusted R^2	0.123	0.010	0.149	0.073
# Obs.	14,768	14,124	15,029	14,443

Panel B. Number of Earnings Call Participants				
	High *N_Participant*		Low *N_Participant*	
	(1)	(2)	(3)	(4)
	$Volume_{it}^{PPT}$	$Return_{it}^{PPT}$	$Volume_{it}^{PPT}$	$Return_{it}^{PPT}$
$Vocal\_Ambiguity_{it}$	-2.1413**	-0.0215***	-0.1771	-0.0069
	(-2.49)	(-3.23)	(-0.18)	(-1.06)
Firm-Level Control	Yes	Yes	Yes	Yes
Script-Level Control	Yes	Yes	Yes	Yes
Year-Fixed Effect	Yes	Yes	Yes	Yes
Industry-Fixed Effect	Yes	Yes	Yes	Yes
Adjusted R^2	0.080	0.051	0.106	0.045
# Obs.	16,897	16,409	12,900	12,159

Table 8. Robustness Tests

*, **, *** represent statistical significance at the 10 percent, 5 percent, and 1 percent levels, respectively.

Panel A reports the regression results using an alternative delivery measure calculated with self-entropy. Panel B reports the regression results using one-day return and trading volume instead of real-time market reactions. Standard errors are clustered by industry. All continuous variables are winsorized at the 1 percent and 99 percent levels. t-values are reported in the parenthesis.

Panel A. Using Alternative Delivery Measure (Based on *N-Gram*)				
	(1)	(2)	(3)	(4)
	$Volume_{it}^{PPT}$	$Return_{it}^{PPT}$	$Volume_{it}^{Q\&A}$	$Return_{it}^{Q\&A}$
$Vocal\_Clarity_{it}$	0.2120**	0.0024***	0.1853***	0.0017***
	(2.26)	(2.83)	(4.82)	(3.60)
Firm-Level Control	Yes	Yes	Yes	Yes
Script-Level Control	Yes	Yes	Yes	Yes
Year-Fixed Effect	Yes	Yes	Yes	Yes
Industry-Fixed Effect	Yes	Yes	Yes	Yes
Adjusted R^2	0.067	0.042	0.298	0.077
# Obs.	29,797	28,568	29,352	27,774

Panel B. Using One-Day Return and Trading Volume		
	(1)	(2)
	$Volume_{it}^{0,1}$	$Return_{it}^{0,1}$
$Vocal\_Ambiguity_{it}^{Call}$	0.1830	0.0012
	(0.72)	(0.36)
Firm-Level Control	Yes	Yes
Script-Level Control	Yes	Yes
Year-Fixed Effect	Yes	Yes
Industry-Fixed Effect	Yes	Yes
Adjusted R^2	0.354	0.640
# Obs.	29,656	29,165

국문초록

이익발표의 음성전달력과 실시간 주식시장 반응

김건우

경영학과 회계학전공

서울대학교 경영대학원

본 연구에서는 기업 이익발표의 음성 전달력과 실시간 주식시장 반응을 분석한다. 기업의 이익발표 중 투자자들은 대본이 없는 상태로 발표를 듣는다. 투자자들은 음성 정보를 처리할 때 문자기반 정보와 비문자기반 정보(음성 등)를 동시에 해석하므로(Zahn, 1973), 이익발표의 음성적 정보와 문자 정보가 모두 투자의사결정에 영향을 줄 것으로 예상된다. 본 연구에서는 딥러닝 기법을 이용하여 미국 이익발표 오디오의 음성 전달력을 측정하고, 문자 정보를 통제한 상태에서 음성 전달이 모호(ambiguous)한 경우 주식시장에서의 반응이 상대적으로 경미함을 보인다. 더불어, 이러한 효과는 기업이 나쁜 실적을 공시할 때 덜 강하게 관측되며 이익발표에 참여하는 애널리스트가 많을수록 더 강하게 관측된다. 이상의 결과는 음성 전달력 측정 방법을 달리 하였을 때에도 유사하게 관측되며, 회귀모형에 이익발표 수준 고정효과와 발표자 수준 고정효과를 모두 포함하였을 때도 강건하게 관측된다. 정리하자면, 본 연구는 정보의 전달 방식이 투자자들의 투자의사결정에 영향을 미침을 시사한다.

주요어: 이익발표, 음성전달력, 정보전달, 실시간 시장반응, 정보처리비용
학번: 2019-26470

Acknowledgment

I am heavily indebted to my academic advisor Bok Baik for his excellent guidance and unwavering support. I appreciate insightful comments and suggestions from Elizabeth Blankespoor, Carlos Corona, Seung-Yeob Han, Carla Hayn, Sung-Gon Jung (discussant), Jung-Koo Kang, Hyejin Kim (discussant), Kuan-Hui Lee, Clive Lennox, Shirley Lu, Charles McClure, Dawn Matsumoto, William Mayew, Tatiana Sandino, Ewa Sletten, Suraj Srinivasan, Eric So, Christopher Stewart, David Park, Charles Wang, Eric Yeung, Jaeho Yoo, Anastasia Zakolyukina and seminar participants at MIT, Seoul National University Sustainable Accounting Seminar, Seoul National University Finance Seminar, 2021 Korea Accounting Information Association Fall Conference, and 2021 Korea Association of Telecommunications Policy Annual Conference. This paper uses earnings call audio and transcript data from S&P Global. I appreciate the research support from S&P Global. Also, I appreciate Artificial Society Inc. for their GPU support. All remaining errors are on my own.