THE UNIVERSITY OF CHICAGO

Predicting Returns with Audio Data

By

Xuran Zeng

June 2022

A paper submitted in partial fulfillment of the requirements for the
Master of Arts degree in the
Master of Arts Program in the Social Sciences

Faculty Advisor: Dacheng Xiu
Preceptor: Christopher Roark

Predicting Returns with Audio Data

Xuran Zeng *

June 2022

Abstract

This article introduces a new sentiment analysis framework that extracts information from audio records to predict stock returns. Unlike the previous scoring approach in finance literature that relies on the vocal emotion analysis software, this approach combines Speech Emotion Recognition and Transfer Learning to deliver customized sentiment scores for individual research purposes. It contains three components: feature extraction, sentence-level transfer learning, and scoring aggregation via penalized likelihood. In the empirical analysis, I study the earning calls corpus and examine the incremental informativeness and the corresponding market reaction. I show that this approach excels at extracting predictive signals and that speech sentiment is a remarkable predictor of abnormal returns. I then extract the text sentiment scores from the corresponding transcript using the BERT model and demonstrate that a trading strategy based on speech generates a Sharpe Ratio that is 2.75 times higher than the text strategy. Moreover, speech contains more information than text and a strategy combining both of them provides a Sharpe Ratio that is twice as large as the pure speech strategy. Eventually, information in speech is incorporated into prices with a delay thus it can be exploited in a real-time trading strategy.

Keywords: Return Predictability, Speech Emotion Recognition, Transfer Learning, Machine Learning

*M.A. Student in MAPSS, The University of Chicago

I would like to thank Dacheng Xiu, Christopher Roark, Christopher Graziul, as well as seminar participants at the University of Chicago for helpful comments and suggestions.

1 Introduction

Stocks return prediction is a remarkable concern for researchers and market practitioners. Traditionally, people have relied on accounting and fundamental data. In recent decades, technology is deeply incorporated into financial studies (surveyed in Giglio et al. 2021). Machine learning makes it possible to use predictive information in complex and unstructured datasets, such as textual data (Ke et al., 2020) and image data (Jiang et al., 2020). Text is the most commonly used data source in finance to study stakeholders' sentiment in news, annual reports, and earnings conference calls transcripts. While the textual and visual analysis in finance is growing increasingly sophisticated, the usage of speech data is in its infancy.

Studies in psychology and linguistics have long documented that the human oral speech conveys extensive information beyond the literal meaning shown in a verbal content (Caffi and Janney, 1994). This suggests that compared to the text, speech communication might be an alternative or even better signal to study sentiment and predict stock returns. In computer science, new techniques have been around for over two decades to deal with audio sentiment measurement. Speech emotion recognition (SER) is a collection of methodologies that process audio signals, extract acoustics features and detect the embedded emotions. Despite the widespread use of SER in psychological assessment, human-computer interaction, and communication evaluation (surveyed in Akçay and Oğuz 2020; Schuller 2018), research that focuses on SER application in finance and economics is limited. In finance, it was not until 2012 that researchers start focusing on sentiment analysis based on nonverbal communication. In early studies, the sentiment scores are frequently computed through a commercial vendor, and then used in asset pricing models for investigating the impact on stock price movement and the firm's future performance.

In this paper, I propose a new method to conduct sentiment analysis on speech data based on SER and cross-corpus transfer learning. This scientific approach establishes a valid measurement of acoustics features and provides meaningful estimates of the speaker's emotional state without relying on the vendor's private "know-how". The intuition of this model is to learn the feature pattern of known sentiment speech and use that knowledge to classify unknown sentiment speech. I abbreviate my method as SETLS (Sentiment Extraction via Transfer Learning and SER) and it consists of three parts. The first step extracts the acoustics features set, eGeMAPS (Eyben et al., 2016), on the sentence level for two datasets: the source domain (labeled records) and the target domain (unlabeled records). The second step conducts the transfer learning-based sentiment classification with a domain confusion loss to learn a representation of the target from the source that minimizes the variance in between (Tzeng et al., 2014). At the end of the second step, I get sentence-level sentiment scores on the target dataset. The third step assigns document-level sentiment scores via penalized maximum likelihood (Ke et al., 2020).

The main virtue of this model is its transparency and flexibility. Compared to the secret algorithms of the commercial product, this model clearly presents the machine learning architecture and the acoustic features based on which the classification is conducted. Therefore, it is much more flexible and there remains a wide-open area to improve the structure or the input. For example, adding textual data and constructing a multi-modality model might improve the accuracy of sentiment classification and empower the returns forecast ability. The implementation of more acoustic features, such as the Teager Energy Operators (Zhou et al., 2001), which are specifically designed to measure stress, enables deception detection during a financial

disclosure presentation.

My empirical analysis revisits from a speech analysis perspective the core aspect of financial markets research: the extent to which sentiments of earnings calls explain and predict observed asset price variation. I analyze the earnings calls dataset proposed by (Qin and Yang, 2019). This dataset contains 571 earnings conference calls of 279 companies in 2017, with a total of 89,843 utterances. Using the dates of the events and the names of the companies, I match the speech with stock data from CRSP and Compustat to analyze the price movements and firms' performance after the disclosure speech. The key feature of my approach is that I learn the sentiment scoring model from the external corpus and based on emotion-related acoustic features, rather than predicting asset prices directly or taking sentiment scores off the shelf.

To translate the machine learning results into economic terms, I study the impacts of speech sentiment scores on firms' returns and evaluate the performance of trading strategies that buy assets with positive speech sentiment and sell assets with negative sentiment. These long-short portfolios perform well on the first two days after the earning conference, while the impact dies out on the following days.

I compare the price impact of speech versus text sentiments by extracting sentiments from earning calls transcripts with the FinBERT (Araci, 2019) model, a language model based on BERT for financial NLP tasks. While long-short strategies based on both forms of sentiment generate significant positive excess returns, the correlation between the two sentiment scores is insignificant. After controlling for the linguistic content in the conference calls, both positive and negative effects exhibited by managers are associated with contemporaneous stock returns. Additionally, a strategy that combines two signals generates a higher Sharpe Ratio than any of the single signal strategies(170% higher than pure speech strategy and 475% higher than pure text strategy).

Furthermore, this framework can be used to investigate the process of price formation. While the speech sentiment information is fully incorporated into prices within 2 days, the text sentiment takes one extra day to be completely assimilated. As a consequence of this high speed of assimilation, I propose intraday and overnight strategies to explore the arbitrage opportunity. Likewise, I study how differences in sentiment information assimilation are associated with firm size and find that price responses to speech sentiment are roughly 28 times as large for smaller firms and that it takes twice as long for sentiment shock about big firms to be fully reflected in prices.

Eventually, I demonstrate the universal adaptability of SETLS by employing an alternative source corpus and classifying speech based on discrete emotions.

This paper adds to the research on nonverbal communication in finance. Most prior work using speech as data for finance and accounting research use a commercial software called Layer Voice Analysis (LVA) to obtain sentiment scores. They do little of vocal acoustic features based sentiment measurement model, and the results obtained from this commercial product are unreliable because it does not publish its algorithm. Early examples are (Mayew and Venkatachalam, 2012; Hobson et al., 2012; Price et al., 2017). Nonverbal communication was first studied by Mayew and Venkatachalam (2012). They measure the emotional states during earnings conference calls using LVA. Then they regress cumulative abnormal returns on the emotion scores. With these regressions, they show that positive and negative effects displayed by CEOs and CFOs are informative about the firm's financial performance in the future. Later, Hobson et al. (2012) further investigated nonverbal signals and find that vocal markers of cognitive dissonance by CEOs during earnings

conference calls are positively associated with the likelihood of irregularity restatements, suggesting that vocal cues may be used to detect intentional deception in corporate disclosures. These two articles are the precedents in the research of speech in the financial field and lay a foundation and paradigm for subsequent research. However, they are criticized for using LVA, an irrelevant technology that cannot measure emotions conveyed by voice (Lacerda, 2012).

Some papers published at computer science conferences adopt machine learning techniques, e.g., BiLSTM or transformer, to predict stocks volatility directly using textual transcripts and acoustics features from audios(Qin and Yang, 2019; Yang et al., 2020; Li et al., 2020). While their models achieve significant and substantial prediction error reduction, they do not explain the linkage between acoustic feature, sentiment, and financial outcomes. This makes them similar to black boxes as the LVA software. On the contrary, the output of my model is sentiment scores, which naturally overcome this deficiency. They have economic meanings because they depict speakers' beliefs and plenty of empirical evidence shows that these beliefs have an impact on asset prices.

Sethuraman et al. 2018 is my closest predecessor and they focus on one acoustics feature, the voice pitch. According to psychological theory, voice pitch decreases when a speaker is more confident, so we expect a negative impact of pitch on stock prices. Using intraday data and a structural equation model, they find that stock prices increase when managers exude dominance in the conversation, as captured through changes in voice pitch. While it draws a reasonable link between voice pitch and confidence, the research scope is limited to one feature and one emotion. In contrast, I propose a classifier that is designed to implement all types of acoustic features (prosodic, spectral, voice quality, energy, etc.).

This is the first paper, to my knowledge, to propose a sentiment classification based on phonetic variables in financial speech. It proposes a processing pipeline capable of transforming audio communications into sentiment scores for further analysis. Thus, it suggests the applicability of a brand new data source for researchers in social science, including finance (e.g., IPO roadshows), economics (e.g., Federal Open Market Committee meetings, interest rate prediction, etc.), and accounting (e.g., earnings calls).

The rest of the paper is organized as follows. In Section 2, I introduce the two components of the SETLS: Speech Emotion Recognition and Transfer Learning. Section 3 reports an empirical analysis of stock-level earnings calls and returns using SETLS.

2 SETLS: A Sentiment Extraction Algorithm based on Transfer Learning and SER

2.1 Speech Emotion Recognition

2.1.1 Emotions

Before implementing the speech sentiment classification model, it is necessary to define emotion. There is no consensus about the definition of emotion, but two models have become common in the SER system: the discrete emotional model and the dimensional emotional model.

Discrete emotion theory is based on the six categories of basic emotions: sadness, happiness, fear, anger,

disgust, and surprise, according to Ekman and Oster (1979). These are inborn and culturally independent emotions. Other emotions are obtained by combining the basic ones.

Dimensional emotion theory is an alternative theory that uses a smaller number of dimensions (valence, arousal, activation, dominance, etc.) to characterize emotions (Russell and Mehrabian, 1977). According to this model, emotions are not independent. Instead, they are analogous to each other in a systematic way. One of the most preferred dimensional models is a two-dimensional model that uses valence and arousal. The valence dimension defines whether an emotion is positive or negative. The arousal dimension describes the strength of the emotion, whether it is high or low.

In this financial setting where managers communicate information to investors about both past and future performance, it is natural to implement the dimensional emotion model and classify utterances into positive or negative. For example, a manager is likely to exhibit positive sentiment if the manager expects positive future firm performance due to private information regarding current outcomes. In contrast, a manager may express negative sentiment when possessing negative private information or experiencing a cognitive dissonance stemming from a deceptive discussion. In this article, I first use a one-dimensional model and classify sentiments into positive and negative. This enables us to examine the impact of positive and negative sentiment on price movements. Later in section 3.6, I explore the impact of discrete emotions and further explain why the dimensional emotion model is more suitable in the context of financial markets.

2.1.2 Feature Selection

Features are an important aspect of SER. An enormous amount of phonic features has been used for SER systems, but there is no generally accepted set of features for precise and distinctive classification. In this article, I choose the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) as the model input. eGeMAPS is a basic standard acoustic parameter set. It contains 25 low-level descriptors, including parameters from three groups: frequency (prosodic), energy (amplitude), and spectral (See detailed description in 1). All features were selected based on two criteria.

First, they have theoretical significance (Scherer, 1986, 2003). Frequency features (pitch, jitter, formant 1,2,3 frequency and bandwidth) are variants of the fundamental frequency $F0$, which is created by the vibrations in the vocal cord. Researchers show that throughout the production of high arousal emotions such as happiness and excitement, mean $F0$ and $F0$ variants increase, while $F0$ contour decreases. In contrast, in the production of low arousal emotions, such as neutral and disappointed, mean $F0$ and its variants decrease, while its contour increases (Frick, 1985). Energy features (shimmer, loudness, HNR) represent the amplitude variation of speech signals over time. Researchers suggest that high arousal emotions yield increased energy while low arousal emotions result in decreased energy (Lin et al., 2012). Spectral features (Alpha Ratio, Hammarberg Index, MFCC, etc.) are constructed by transforming the time domain signal into a frequency domain signal using the Fourier transform. When sound is produced by a person, it is filtered by the shape of the vocal tract. Characteristics of the vocal tract are represented in the frequency domain. Therefore, the transformation from time to frequency enables spectral features to approximate the human auditory system's response more closely and thus result in a more accurate representation of the sound (Scherer, 2003).

Second, they have been frequently used and proven effective in previous research. Examples include

Table 1: eGeMAPS features description

Group	Features	Description
Frequency	Pitch	logarithmic F0 on a semitone frequency scale starting at 27.5 Hz (semitone 0)
	Jitter	deviations in individual consecutive F0 period lengths
	Formant 1, 2, and 3 frequency	centre frequency of first, second, and third formant
	Formant 1, 2, and 3 bandwidth	bandwidth of three formants
Amplitude	Shimmer	difference of the peak amplitudes of consecutive F0 periods
	Loudness	estimate of perceived signal intensity from an auditory spectrum
	Harmonics-to-noise ratio (HNR)	relation of energy in harmonic components to energy in noise-like components
Spectral	Alpha Ratio	ratio of the summed energy from 50-1000 Hz and 1-5 kHz
	Hammarberg Index	ratio of the strongest energy peak in the 0-2 kHz region to the strongest peak in the 2–5 kHz region
	Spectral Slope 0-500 Hz and 500-1500 Hz	linear regression slope of the logarithmic power spectrum within the two given bands
	Formant 1, 2, and 3 relative energy	as well as the ratio of the energy of the spectral harmonic peak at the first, second, third formant's centre frequency to the energy of the spectral peak at F0
	Harmonic difference H1-H2	ratio of energy of the first F0 harmonic (H1) to the energy of the second F0 harmonic (H2)
	Harmonic difference H1-A3	ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3)
	MFCC 1-4	Mel-Frequency Cepstral Coefficients 1-4
	Spectral flux	difference of the spectra of two consecutive frames

Frequency features (Schuller et al., 2003; Rao et al., 2013), Energy features (Li et al., 2007; Zhang, 2008), and Spectral features (Kuchibhotla et al., 2014; Wong and Sridharan, 2001)

The features extraction implementation of eGeMAPS is publicly available with the openSMILE toolkit.

2.2 Transfer Learning

A bottleneck of studies on speech sentiment measurement is the lack of labeled corpus. Traditional machine learning-based sentiment classifiers are characterized by training data and testing data with the same input feature space and the same data distribution. However, the vocal features of more well-defined corpora and this understudied earning calls audio data are different. This implies the poor predictive performance of models trained on labeled corpora then applied to an unlabeled corpus. Transfer learning can solve this issue. Schuller et al. (2010) implemented several normalization strategies for transfer learning and showed results employing six standard databases in a cross-corpora evaluation experiment. Deng et al. (2013) developed a sparse autoencoder-based feature transfer learning model and conducted an experiment on six standard databases. Zhang et al. (2017) put forward a cross-corpus multi-task learning model and evaluated model performance on EmoDB, eNTERFACE, VAM, and AVEC datasets. All these models have a significant improvement in performance, but they all focus on standard datasets. In practice, naturalistic audio corpora are more complicated and, given their prevalence in real-world applications, worth studying. Additionally, these methods depend on labeled testing data, even though only a small amount is needed. In the case of financial audio, there are no labels at all. Therefore, these methods cannot classify sentiment contained in earning calls.

The SETLS model is intended to transfer the learned representations in the labeled dataset to a completely unlabeled dataset while avoiding overfitting.

2.2.1 Overall Architecture

To successfully transfer the knowledge, I hope to capture the invariant in the two datasets and minimize the discrepancy. Therefore, the model is devised into two parts. Firstly, labeled data is used to train the model so that it can classify data into different categories given the acoustics features (on the left). The second part (on the right) is aimed at minimizing the distance between the two domains. The distance is measured by the Maximum Mean Discrepancy (MMD Tzeng et al. 2014) calculated on the representations output from the adaption layer. The losses from the two parts are combined as the final loss for optimizing the parameters of the entire model. Figure 1 exhibits the overall architecture.

The adaptation layer on the left is intended for accurate sentiment classification of labeled samples. In this model, I use a fully connected neural network followed by a softmax layer to compute the probabilities of the labels for each sample.

$$p_{x,c} = softmax(W_O(x_{GeMAPS}) + h_O) \tag{1}$$

. The x is restricted to labeled data. The W, h are trainable parameters.

The adaption layer on the right has a lower dimension, since it should regularize the representation to be invariant to the source and target domains and prevent overfitting to some unrepresentative noise in the two domains.

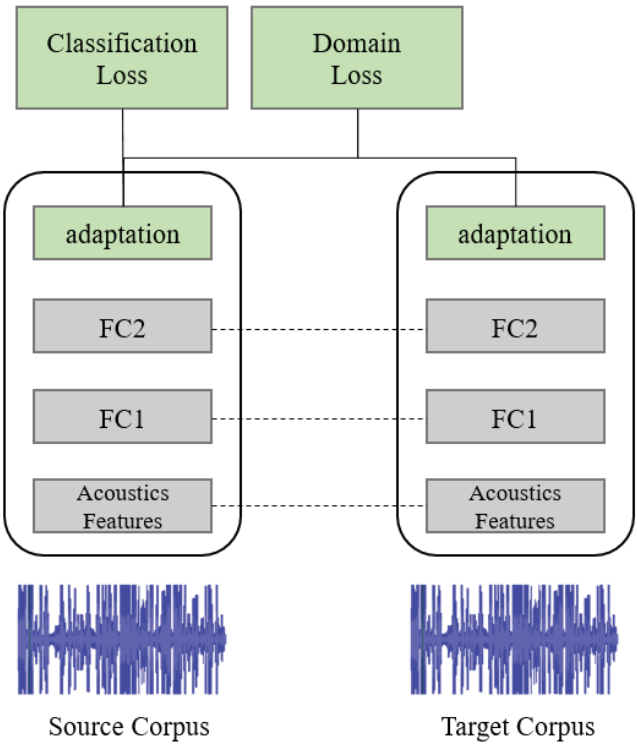

Figure 1: Model framework.

$$\phi(x) = AdaptionLayer(x_{GeMAPS}) \tag{2}$$

The x could be either an source sample or a target sample.

For this 2-category classification, I use the cross entropy function to calculate the classification loss. Specifically, I calculate a separate loss for each class per observation and sum the loss:

$$L_C(X_L, y) = -\sum_{x \in X_L} \sum_{c=1}^{|C|} y_{x,c} \log(p_{x,c}) \tag{3}$$

, where $p_{x,c}$ is the predicted probability that the observation x has the label c.

For the domain loss, I use the maximum mean discrepancy to measure the distance between the source dataset and the target dataset, which could be expressed as the following equation.

$$MMD(X_S, X_T) = ||\frac{1}{|X_S|} \sum_{x_s \in X_S} \phi(x_s) - \frac{1}{|X_T|} \sum_{x_t \in X_T} \phi(x_t)|| \tag{4}$$

In this model, the representation $\phi(\cdot)$ is obtained by passing the features through adaption layer.

2.2.2 Source Domain

For the source domain, I select the Speech Under Simulated and Actual Stress (SUSAS Hansen et al. 1997) database, which is widely used for speech emotion recognition. SUSAS contains recordings of 32 speakers over five domains, including acted, elicited, and natural speech. Utterances in acted speech are recorded by professional actors in sound-proof studios. Elicited speeches are created by placing speakers in a simulated emotional situation. Natural speeches are obtained from real-world situations. I choose the

recordings from actual speech to minimize the difference between source and target corpus. Each utterance was annotated using five categories: high stress, medium stress, screaming, fear, and neutral. To obtain a dimensional sentiment classification, I map the emotions for two binary classification tasks: arousal and valence, following the general process of studies on this corpus (Schuller et al., 2010). Low arousal sentiment is neutral. High arousal sentiments are high stress, medium stress, screaming, and fear. Negative valence sentiments are high stress, screaming, and fear. Positive valence sentiments are medium stress and neutral. A total of 16,000 utterances are used.

With regards to the discrete emotional model, I use the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS Livingstone and Russo 2018) as the source database. This is an English language emotion dataset that is widely used for emotional song and speech recognition. The dataset consists of 24 actors (12 male and 12 female) to record eight different emotions: anger, calm, happy, sad, surprise, disgust, neutral and fearful. A total of 7,356 utterances are utilized for training.

3 Empirical Analysis

3.1 Data and Pre-processing

For empirical analysis, I use a sample of 571 firms' quarterly earning conferences collected by Qin and Yang (2019). It contains 89,843 unique utterances from January 17, 2017 to December 21, 2017, covering 279 separate SP 500 firms appearing an average of 2.05 times. With a maximum (minimum) of 4(1) observations per firm and a standard deviation of 0.97, no single firm comprises more than 0.70 % of the sample. Figure 2 plots the average number of earning calls per day over a year. It shows the quarterly earnings season effects around February, May, August, and November.

Using the company name, day, and permno, I match each call with the firm's market capitalization, book-to-market ratio, and open-to-open return from CRSP and Compustat.

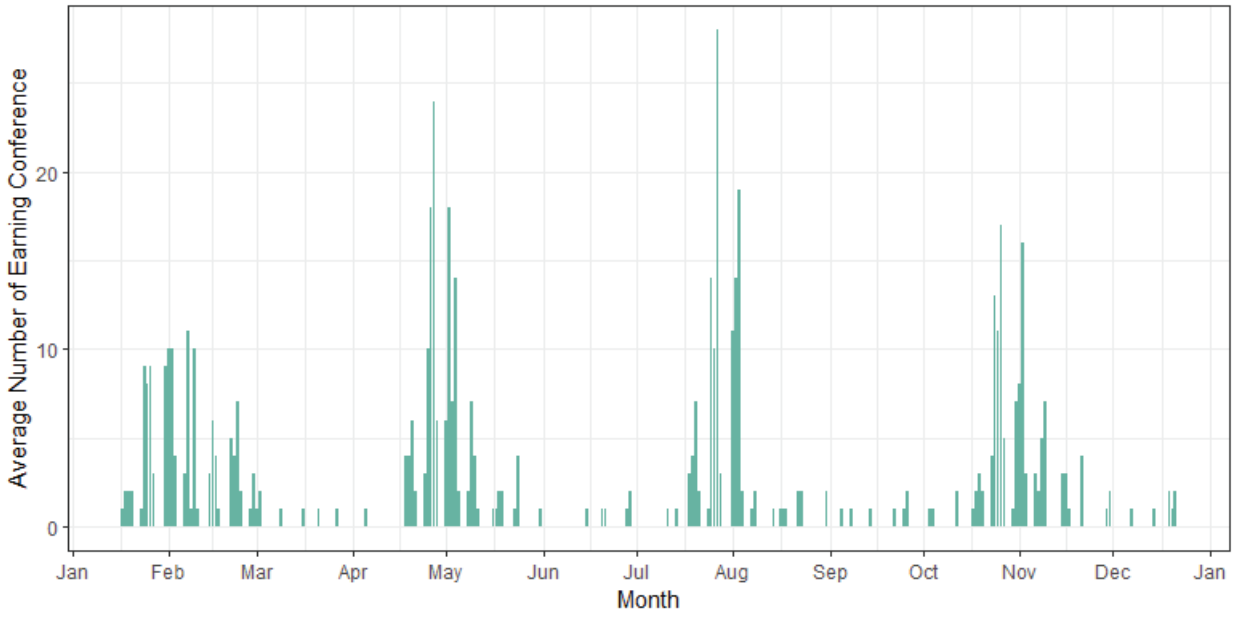

Figure 2: Average Audio Counts

To gauge the investor response to nonverbal emotion cues, I calculate cumulative abnormal returns (CARs) during a one-day initial reaction window and check the correctness of the response using a subsequent five-day period. To investigate the long-term impact, I use a subsequent 180-day period to calculate CAR(2,180), which approximately measures the return movement between two earnings disclosure. Besides, since the first day is not tradable, investors are also interested in the cumulative returns on the second day captured by CAR(1,2). CAR(0,1) is calculated as the difference between each firm's daily return and the return on the SP 500 index and then summed from the day of the conference call (trading-day t=0) to one day after (trading day t=1). I further check alternative windows using CAR(1,2), CAR(2,5), and CAR(2,180), where the specific trading-day ranges are indicated within the parentheses.

Figure 3 describes the earning conference timeline and the trading activities. There is not any strict standard on when a company holds the earning conference. Some companies choose to report before the market opens, while others report after standard market hours. Therefore, I use the open-to-open returns to capture the price movement. For earning calls that occur on day 0, I build positions at the market opening on day 1, and rebalance at the next market opening, holding the positions of the portfolio within the day. This portfolio is defined as a day +1 portfolio. Similarly, I can define day 0, day -1, etc. In section 3.4, I build positions according to the exact timing of earning calls and discuss the returns of intraday and overnight portfolios.

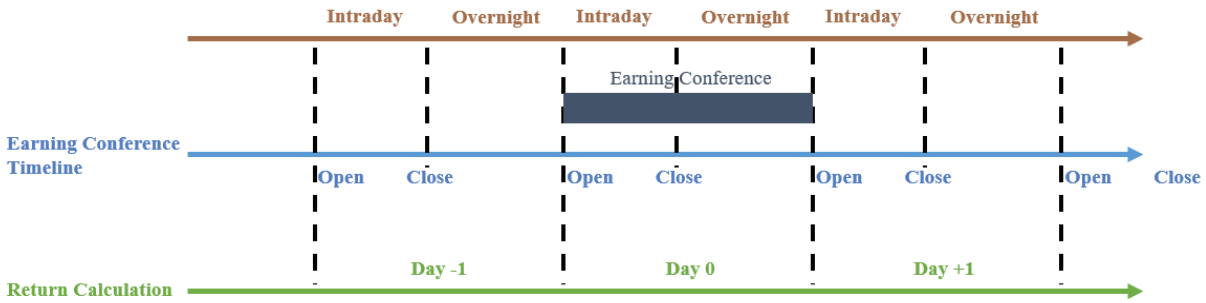

Figure 3: Event Timeline

Size is the log of market capitalization 20 trading days prior to the conference call. BM is the ratio of book equity to market equity at the end of the previous quarter. Following Fama & French, these two factors were employed as the control variables.

I estimate the sentiment scores via SETLS using the SUSAS corpus as the source domain and the earning calls corpus as the target domain. For each sentence, I generate a 25-dimension audio vector to represent vocal features. Then I train the model and predict the positive or negative sentiment for each sentence. Eventually, the maximum likelihood estimator is used to obtain sentiment scores at a document level.

Table 2 provides the descriptive statistics and unconditional correlation for the Sentiment Scores. The unconditional correlations in Table 2 suggest that investors initially react positively to executive emotion (coefficient of 0.13 and 0.08). In the following days, the relation slightly reverses with a coefficient of -0.04.

Table 2: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max	Sentiment Correlation
Sentiment Scores	536	0.694	0.128	0.180	1.000	1.000
CAR(0,1)	536	−0.001	0.038	−0.206	0.280	0.130
CAR(1,2)	536	−0.001	0.038	−0.225	0.155	0.077
CAR(2,5)	536	−0.002	0.026	−0.151	0.166	−0.036
CAR(2,180)	536	−0.015	0.186	−0.835	1.489	0.041
LNME	536	10.139	1.188	1.818	13.264	0.007
BM	536	0.291	0.280	−0.228	2.208	−0.076

3.2 Impact on returns

To investigate the impact of speech sentiments on returns, I sort firms into 5 portfolios based on the sentiment scores and examine portfolio-level risk and returns. The top 20% of firms with the highest sentiment scores are divided into group 5, and the bottom 20% of firms with the lowest sentiment scores are divided into group 1. Each group in the middle is the stocks in descending order by 20%. I then document the statistics of stock performance in each group. Finally, I construct long-short portfolios by longing the top group and shorting the bottom group. With reference to investment research, I use both equal-weighted averaging strategy and value-weighted strategy (weighted by the log of market value). Equal-weighted is a simple and robust means of assessing the predictive power of sentiment. Value-weighted gives more weight to firms with large market capitalization.

Table 3 Panel A indicates that initial returns monotonically increase from the bottom group (-1.02%) to the top group (0.27%). Both the equal-weighted and value-weighted long-short portfolios have significantly positive returns (1.30%, 1.19%) with t-statistics higher than 2. Panel B demonstrates that the portfolio returns during the subsequent period continue to increase moving from negative (-0.42%) to positive sentiment(0.33%), though the coefficients are less significant than the previous day. The long-short difference is 0.75% for equal-weighted and value-weighted portfolios. After the first two days, Panel C illustrates that the speech sentiment has hardly any impact on returns. In the long run, Panel D suggests that the returns movement triggered by executive sentiment diminishes. As the sentiment scores increase, neither the average return nor the unit risk-return rate shows monotonicity. The T statistics of the long-short portfolio are 0.521 and 0.43, which fail the test. Mayew and Venkatachalam (2012) finds that negative affect is related to cumulative abnormal returns over the subsequent 180 trading days following the earnings conference call using data in 2007. In Price et al. (2017) as well as this study, no long-term effect is found, suggesting a higher market efficiency.

Table 3: Group Statistics: Speech Sentiment

Panel A: CAR(0,1)

	Portfolios Sorted on Sentiment Scores					Equal	Value
	1	2	3	4	5	5 minus 1	5 minus 1
Excess Return	-1.02%	-0.35%	-0.15%	0.08%	0.27%	1.30%	1.19%
Standard Deviation	4.27%	3.05%	3.40%	2.68%	3.59%	5.70%	5.63%
Sharpe Ratio	-0.24	-0.11	-0.04	0.03	0.08	0.23	0.21
TSTAT	-2.37	-1.14	-0.43	0.30	0.76	2.26	2.10
PVALUE	0.02	0.26	0.67	0.76	0.45	0.03	0.04
Amount of Factor 0.48	0.69	0.73	0.76	0.81			

Panel B: CAR(1,2)

	Portfolios Sorted on Sentiment Scores					Equal	Value
	1	2	3	4	5	5 minus 1	5 minus 1
Excess Return	-0.42%	-0.14%	-0.10%	0.05%	0.33%	0.75%	0.75%
Standard Deviation	4.16%	3.10%	2.76%	4.10%	4.01%	5.82%	5.79%
Sharpe Ratio	-0.10	-0.05	-0.04	0.01	0.08	0.13	0.13
TSTAT	-1.00	-0.45	-0.37	0.12	0.81	1.27	1.29
PVALUE	0.32	0.65	0.71	0.91	0.42	0.21	0.20
Amount of Factor	0.48	0.69	0.73	0.76	0.81		

Panel C: CAR(2,5)

	Portfolios Sorted on Sentiment Scores					Equal	Value
	1	2	3	4	5	5 minus 1	5 minus 1
Excess Return	-0.16%	-0.32%	-0.24%	-0.37%	-0.01%	0.15%	0.24%
Standard Deviation	2.93%	2.12%	2.42%	2.74%	2.90%	4.00%	3.94%
Sharpe Ratio	-0.05	-0.15	-0.10	-0.14	0.00	0.04	0.06
TSTAT	-0.55	-1.50	-0.97	-1.34	-0.03	0.38	0.61
PVALUE	0.59	0.14	0.33	0.18	0.98	0.70	0.54
Amount of Factor	0.48	0.69	0.73	0.76	0.81		

Panel D: CAR(2,180)

	Portfolios Sorted on Sentiment Scores					Equal	Value
	1	2	3	4	5	5 minus 1	5 minus 1
Excess Return	-2.16%	-4.17%	-2.71%	1.83%	-0.59%	1.56%	1.24%
Standard Deviation	16.36%	17.75%	17.52%	14.77%	23.15%	29.74%	28.62%
Sharpe Ratio	-0.13	-0.23	-0.15	0.12	-0.03	0.05	0.04
TSTAT	-1.30	-2.33	-1.53	1.23	-0.25	0.52	0.43
PVALUE	0.20	0.02	0.13	0.22	0.80	0.60	0.67
Amount of Factor	0.48	0.69	0.73	0.76	0.81		

To verify the univariate inferences with a greater sense of confidence, I utilize all the observations instead of only the top and bottom 20%, and regress CAR(0,1), CAR(1,2), CAR(2,5) and CAR(2,180) on the sentiment scores individually. Table 4 reports the results. The first two columns demonstrate that the coefficients for sentiment scores are significantly positive (0.038 and 0.023) during the initial reaction window. The third column shows a mean revert tendency thereafter, though not significant. And the last column shows an insignificant impact in the long run.

Table 4: CARs Regression on Sentiment Scores

	Dependent variable:			
	CAR(0,1)	CAR(1,2)	CAR(2,5)	CAR(2,180)
	(1)	(2)	(3)	(4)
Sentiment Scores	0.038***	0.023*	−0.007	0.060
	(0.013)	(0.013)	(0.009)	(0.063)
Constant	−0.027***	−0.017*	0.003	−0.056
	(0.009)	(0.009)	(0.006)	(0.044)
Observations	536	536	536	536
R^2	0.017	0.006	0.001	0.002
Adjusted R^2	0.015	0.004	−0.001	−0.0002
Residual Std. Error (df = 534)	0.037	0.038	0.026	0.186
F Statistic (df = 1; 534)	9.260***	3.182*	0.703	0.903
Note:		*p<0.1; **p<0.05; ***p<0.01		

To further investigate the observed relationship in the presence of other factors known to affect stock returns, I regress CAR(0,1) on two control variables, size and value factors. Table 5 shows that controlling size and value factors, the coefficients of the sentiment scores are still significantly positive. The factor loadings on size are negative, suggesting the potentially small size effect. In contrast, the loadings on the book-to-market ratio are positive, which is consistent with the value effect. Both of them are not significant. I will analyze the stock heterogeneity in section 3.5.

Next, I analyze trading strategies that trade in response to speech sentiment with various time delays. These strategies long the top one-third with the most positive sentiment scores and short the bottom one-third with the most negative sentiment scores (Equal-weighted). I focus on one-day open-to-open returns initiated anywhere from one to 10 days following the disclosure. Figure 4 visualizes the average returns in percentage per day with shaded 95% confidence intervals. It shows the long-short portfolio as well as the long and short sides separately. The earning calls take place at time t=0. The return of time t=1 is the change from the event day's open price to the next day's open price. As I defined in section 3.1, this is the day 0 return. Table 6 reports average returns in the first three days.

Table 5: CARs Regression on Sentiment Scores and Controls

	Dependent variable:		
	CAR(0,1)		
	(1)	(2)	(3)
Sentiment Scores	0.038***	0.040***	0.040***
	(0.013)	(0.013)	(0.013)
LNME	−0.001		−0.0003
	(0.001)		(0.001)
BM		0.008	0.008
		(0.006)	(0.006)
Constant	−0.020	−0.031***	−0.028
	(0.016)	(0.009)	(0.017)
Observations	536	536	536
R^2	0.018	0.021	0.021
Adjusted R^2	0.014	0.017	0.015
Residual Std. Error	0.037 (df = 533)	0.037 (df = 533)	0.037 (df = 532)
F Statistic	4.784*** (df = 2; 533)	5.689*** (df = 2; 533)	3.799** (df = 3; 532)

Note: *p<0.1; **p<0.05; ***p<0.01

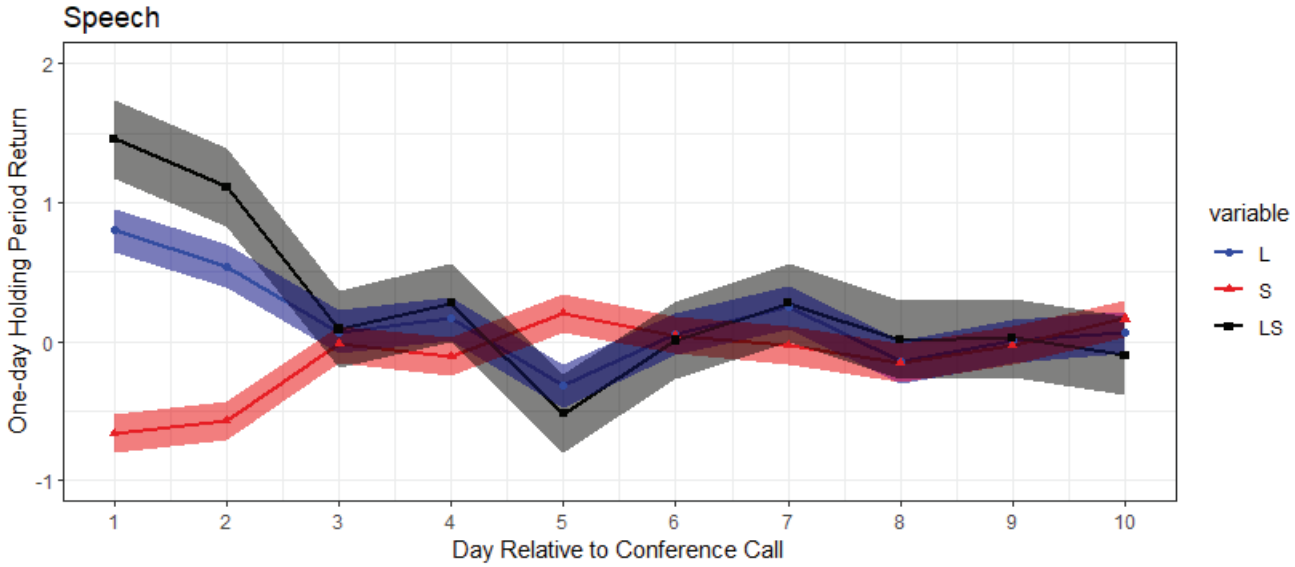

Figure 4: 1D Holding Period Return

Table 6: Performance of Daily Speech Sentiment Portfolios

	Day 0			Day 1			Day 2		
	L	S	LS	L	S	LS	L	S	LS
Average Return	0.81%	0.66%	1.46%	0.51%	0.60%	1.08%	-0.07%	-0.04%	-0.13%
Standard Deviation	4.19%	3.90%	5.76%	4.20%	3.96%	5.70%	2.24%	2.44%	3.11%
Sharpe Ratio	0.19	0.17	0.25	0.12	0.15	0.19	-0.03	-0.02	-0.04
TSTAT	2.63	2.21	3.30	1.62	1.96	2.46	-0.41	-0.22	-0.55
PVALUE	0.01	0.03	0.00	0.11	0.05	0.01	0.68	0.83	0.58

This figure demonstrates the speed of information assimilation. On the third day, one-day holding period returns of the three portfolios become zero, which means that the speech sentiment information is completely incorporated into prices within 3 days.

During the report day, the long side increases by 0.81 % and the short side decreases by 0.66%. Therefore, the long-short portfolio generates a return of 1.46 %. The t-statistics for the three portfolios are all highly significant. Though the return of day 0 strategy is appealing, this is not an implementable strategy because the timing of earning conference would not generally allow a trader to take a position and exploit the return at time t=1. In practice, what investors can do is to build a long-short portfolio at the market open of the day after the earning conferences. That is the day 1 strategy. Though less profitable than the day 0 strategy, this one can earn a return of 0.51% from the long side and 0.60% from the short side, which implies a total return of 1.08% for the long-short portfolio, and the returns are significant. On the third day, the speech sentiment information is entirely reflected in the prices so none of the three strategies could generate significant returns thereafter.

3.3 Textual Sentiment versus Speech Sentiment

In this section, I utilize the most commonly studied unstructured data, earning calls transcripts, to verify whether speech contained additional sentiment information other than the text.

To begin with, I use the transcripts of the earning calls collected by Qin and Yang (2019) and the FinBERT model proposed by Araci (2019) to extract the sentiment of each sentence. Then the penalized maximum likelihood estimation (Ke et al., 2020) is conducted to aggregate terms into an article-level score.

Table 7 reports the outcomes of regression on speech and text sentiment and controlled variables. Controlling the size and value factors, the coefficients of both speech and text sentiments are significantly positive. The R-squared of the third column is 0.029, which is greater than the first two columns (0.021 and 0.009). This indicates that the combination of speech and text sentiments has incremental explanatory power for the post-earnings-announcement drift.

Then, I form the 9 speech-text sentiment stock portfolios following Fama and French (1993). First of all, I allocate stocks into three speech sentiment quantiles and three text sentiment quantiles. Then the 9 speech-text sentiment portfolios are formed as the intersection.

Table 8 shows that the 9 stock portfolios produce a wide range of average excess returns from -1.41%

Table 7: Regression on Speech and Text Sentiment Scores

	Dependent variable:		
	CAR01		
	(1)	(2)	(3)
speech	0.040***		0.042***
	(0.013)		(0.013)
text		0.045*	0.050**
		(0.024)	(0.023)
LNME	−0.0003	−0.0004	−0.0003
	(0.001)	(0.001)	(0.001)
BM	0.008	0.008	0.010*
	(0.006)	(0.006)	(0.006)
Constant	−0.028	−0.029	−0.063***
	(0.017)	(0.022)	(0.024)
Observations	536	536	536
R^2	0.021	0.009	0.029
Adjusted R^2	0.015	0.004	0.022
Residual Std. Error	0.037 (df = 532)	0.038 (df = 532)	0.037 (df = 531)
F Statistic	3.799** (df = 3; 532)	1.699 (df = 3; 532)	4.023*** (df = 4; 531)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 8: Excess returns on 9 stock portfolios formed on speech and text sentiment

	Speech quantiles					
Text	Negative	Medium	Positive	Negative	Medium	Positive
quantile		Means		Standard deviations		
Negative	-1.41%***	-0.12%	0.67%	3.05%	3.39%	4.79%
Medium	-0.86%	-0.24%	0.39%	4.37%	2.69%	3.75%
Positive	-0.11%	-0.32%	0.96%*	4.05%	2.62%	4.38%
	t-statistics for means			p-values for means		
Negative	-3.24	-0.30	1.04	0.00	0.77	0.30
Medium	-1.47	-0.72	0.83	0.15	0.47	0.41
Positive	-0.21	-0.91	1.71	0.83	0.37	0.09

to 0.96%. The portfolios also confirm that there is a positive relation between speech/text sentiment and average return. The relation between average return and speech sentiment scores is more consistent. In every text quantile, average returns tend to increase from the negative- to the positive-speech sentiment portfolios. The portfolio with positive speech and text sentiment has the highest average return, while the portfolio with negative speech and text sentiment has the lowest average return. Therefore, I compare three groups of strategies: pure speech, pure text, and the combination of speech and text. I report the long and short sides separately, as well as the overall long-short strategy performance. Since Day 0 strategy is significant but not tradable, I also report the Day 1 strategy.

Three basic facts emerge from the one-day holding period return of pure text, pure speech and text & speech strategies demonstrated in Figure 5, 6, 7, and Table 9, 10, 11.

Firstly, the combined strategy performs better than the text strategy and the speech strategy. On Day 0, the Sharpe Ratio of the combined strategy reaches 0.38, which is 172 % greater than the speech strategy and 475 % greater than the text strategy. On Day 1, the Sharpe Ratio of the combined strategy is 0.30, which is twice as large as the speech and the text strategy. This finding is supported by the communication theory. In face-to-face communication, only a small fraction of the message regarding the emotional state is conveyed by the verbal content according to Mehrabian et al. (1971). A remarkable component of the message is contained in vocal attributes such as volume, accent, speed, and intonation. Consequently, the speech sentiment extracted based on acoustic features add up to the informativeness of communication and this incremental information gives rise to greater predictability for returns.

Apart from that, Figure 5 and Table 9 both demonstrate that the impact of text sentiment is more significant during the second day after the report day than during the report day. This is consistent with early studies about delayed response. Bernard and Thomas (1989) argue that the market fails to immediately incorporate the information in the earnings textual signal. Abarbanell and Bernard (1992) indicate that the delayed drift may be rooted in the failure of market participants to fully understand the information content of current earnings. Such information includes both the numerical value of the earnings and statements that accompany the earnings announcement. These statements add to the difficulty to understand firm performance and thus lead to the delay in price responses. While the sentiment information in the text is impounded into prices with a delay while the impact of speech sentiment is synchronous and timely.

Besides, Figure 6 shows that speech sentiment information is essentially fully incorporated into prices by the start of Day +3. While Figure 5 illustrates that the text sentiment takes one extra day and achieves full price incorporation by the start of Day +4. How exactly the brain extracts the sentiment of textual words and speech utterances and the relative speed to deal with two signals are largely unknown to researchers in cognitive science. The evidence here suggests that investors and the market take less time to process speech sentiment than text sentiment. Meanwhile, it shows that SETLS is able to identify more complex speech sentiment information contained in the earning calls that investors cannot fully act on using the text information within the first day of trading. Thus, the trading strategy based on speech sentiment is profitable.

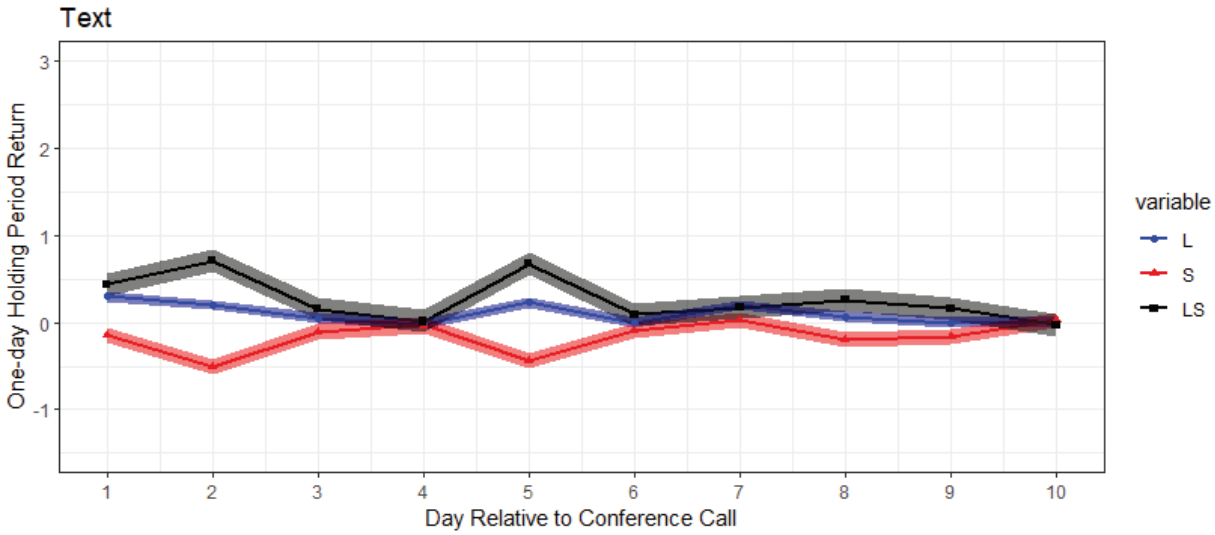

Figure 5: Price Responses of Portfolios Formed on Text Sentiment

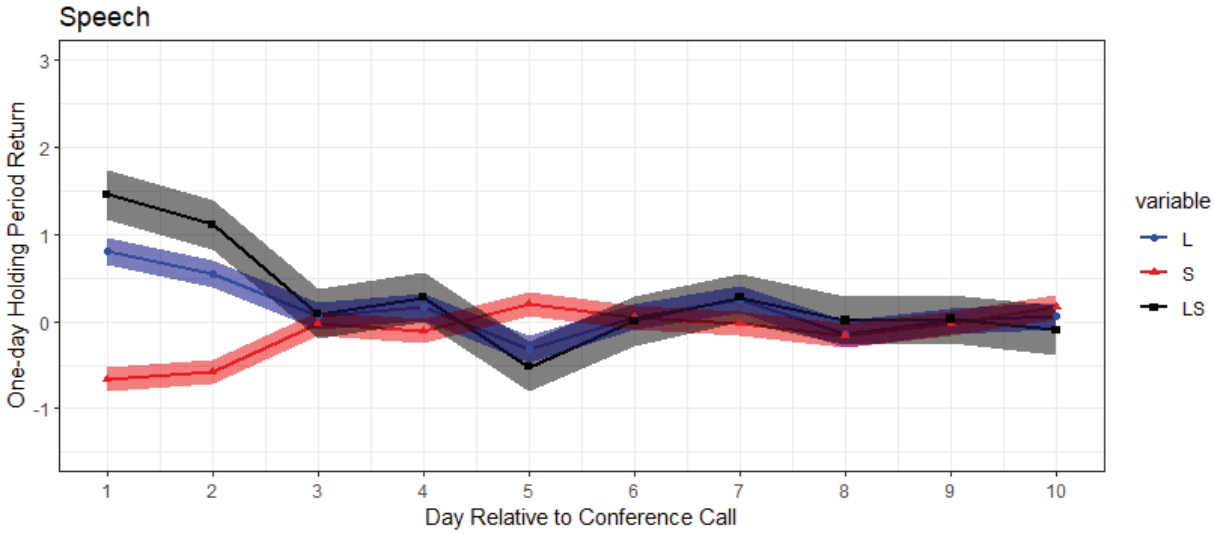

Figure 6: Price Responses of Portfolios Formed on Speech Sentiment

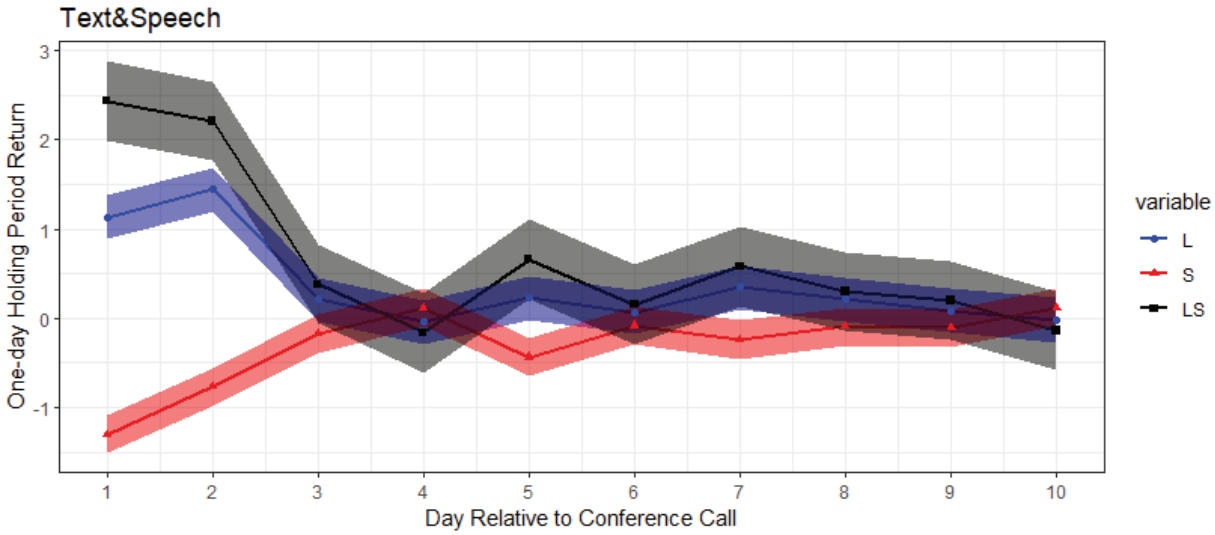

Figure 7: Price Responses of Portfolios Formed on Speech and Text Sentiment

Table 9: Price Responses of Portfolios Formed on Text Sentiment

	Day 0			Day 1		
	L	S	LS	L	S	LS
Excess Return	0.19%	0.23%	0.41%	0.10%	0.57%**	0.74%*
Standard Deviation	3.81%	3.87%	5.41%	3.77%	3.51%	5.34%
Sharpe Ratio	0.05	0.06	0.08	0.03	0.16	0.14
TSTAT	0.66	0.80	0.99	0.36	2.13	1.82
PVALUE	0.51	0.43	0.32	0.72	0.03	0.07

Table 10: Price Responses of Portfolios Formed on Speech Sentiment

	Day 0			Day 1		
	L	S	LS	L	S	LS
Excess Return	0.67%**	0.74%**	1.19%***	0.40%	0.60%**	0.84%*
Standard Deviation	4.29%	3.91%	5.32%	4.15%	3.87%	5.71%
Sharpe Ratio	0.16	0.19	0.22	0.10	0.15	0.15
TSTAT	2.09	2.45	2.88	1.29	2.00	1.91
PVALUE	0.04	0.02	0.00	0.20	0.05	0.06

Table 11: Price Responses of Portfolios Formed on Speech and Text Sentiment

	Day 0			Day 1		
	L	S	LS	L	S	LS
Excess Return	0.96%*	1.41%***	1.83%***	1.22%**	0.85%	1.76%**
Standard Deviation	4.38%	3.05%	4.79%	3.94%	4.01%	5.94%
Sharpe Ratio	0.22	0.46	0.38	0.31	0.21	0.30
TSTAT	1.71	3.24	2.67	2.42	1.47	2.07
PVALUE	0.09	0.00	0.01	0.02	0.15	0.04

3.4 Overnight Strategy versus Intraday Strategy

Since the market response to speech sentiment is at a high speed, the open-to-open strategy is delayed and can miss the best timing. For reports that happen before the market open of Day 0, if investors build the position at the open of Day 1, the impact of speech sentiment might be already incorporated into prices and therefore the strategy performance is degraded. Hence in this section, I will discuss the overnight and intraday strategies and whether they are more profitable than the open-to-open strategy.

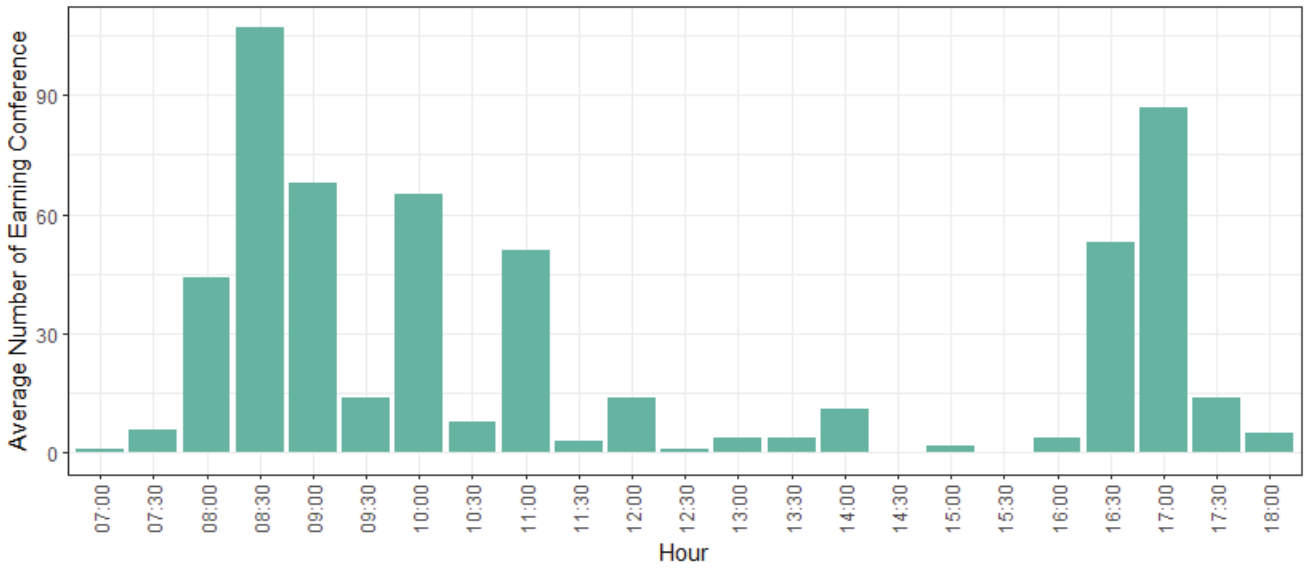

Figure 8: Average Number of Earning Conference by Clock Time

Figure 8 demonstrates the intraday distribution of earning conferences. To begin with, I divide all earning calls audio records into three groups: before the market opens, during the trading hours, and after the market close. Normally, an earning call lasts for one hour. Consequently, calls that happen before 8:30 ET are classified as the 'before' group (158 observations), calls that happen from 8:30 EST to 15:00 ET are defined as the 'intragroup (245 observations), and the rest belongs to the 'after' group (163 observations). For each group, I compare three strategies: overnight, intraday, and holding for an entire day. For 'before' group (Figure 9), the three holding periods are Day 0 open to Day 1 open (entire day), Day 0 open to Day 0 close (intraday), and Day 0 close to Day 1 open (overnight). For 'intragroup (Figure 10), the three holding periods are Day 0 close to Day 1 close (entire day), Day 0 close to Day 1 open (overnight), and Day 1 open to Day 1 close (intraday). For the 'after' group (Figure 11), the three holding periods are Day 1 open to Day 2 open (entire day), Day 1 open to Day 1 close (intraday), and Day 1 close to Day 2 open (overnight). Table 12, 13, and 14 report the strategy performance of each holding period.

Recall that Table 10 shows that the highest Sharpe Ratio of the Day 0 strategy (infeasible) is 0.22 and of Day 1 strategy (feasible) is 0.15, refining the holding periods remarkably improves the returns and Sharpe Ratios. For calls that occur before 8:30 (Table 12), longing the most positive stocks at the market open and rebalancing at the market close generates an excess return of 70 basis points and a Sharpe Ratio of 0.24, which is higher than the Day 0 and Day 1 strategies. Holding the position from open to open can obtain a 74 bp return and 0.26 Sharpe Ratio for the long portfolio and a 108 bp return and 0.25 Sharpe Ratio for

the long-short portfolio, significantly exceeding the 0.15 Sharpe Ratio of the Day 1 strategy. The portfolios that build at the market close still achieve positive payoff but are less significant, which shows that the market is incorporating speech sentiment information into prices and the arbitrage gap is narrowing.

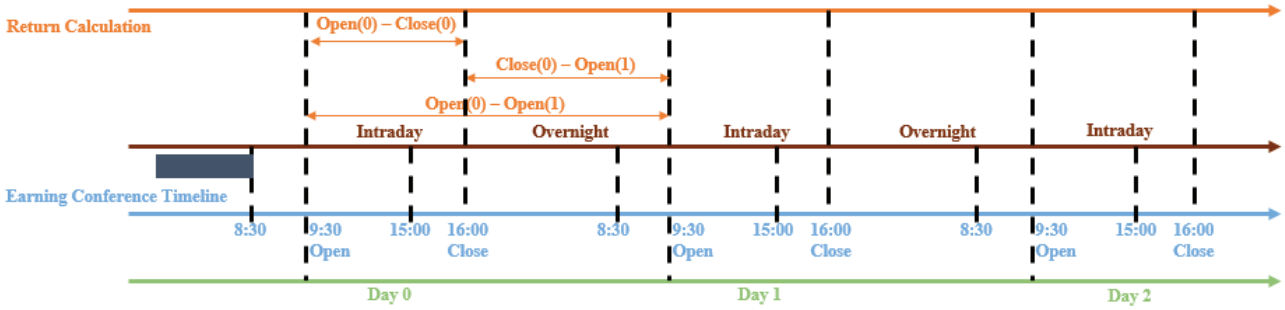

Figure 9: Earning Conference Timeline (before 8:30)

Table 12: Earning Conference before 8:30

	Day 0 open to Day 1 open			Day 0 open to Day 0 close			Day 0 close to Day 1 open		
	L	S	LS	L	S	LS	L	S	LS
Excess Return	0.74% *	0.42%	1.08% *	0.70%	0.25%	0.87%	0.00%	0.12%	0.22%
Standard Deviation	2.89%	3.11%	4.34%	2.96%	2.76%	4.14%	0.88%	0.86%	1.18%
Sharpe Ratio	0.26	0.13	0.25	0.24	0.09	0.21	0.10	0.14	0.19
TSTAT	1.79	0.97	1.74	1.65	0.67	1.47	0.70	1.01	1.30
PVALUE	0.08	0.34	0.09	0.11	0.51	0.15	0.49	0.32	0.20

For earning conferences that take place between the market open and market close (Table 13), holding the position from close to close is less profitable than holding it overnight for the long-only and long-short portfolios. The long-short portfolio earns a Sharpe Ratio of 0.17 overnight, exceeding the feasible Sharpe Ratio of the Day 1 strategy. If investors build the position at the next market open and hold it within the next day, the excess return for both long-only and long-short portfolios becomes negative. This is due in part to the fact that investors tend to overreact to speech sentiment when it first came out, and the prices mean-revert during the next day.

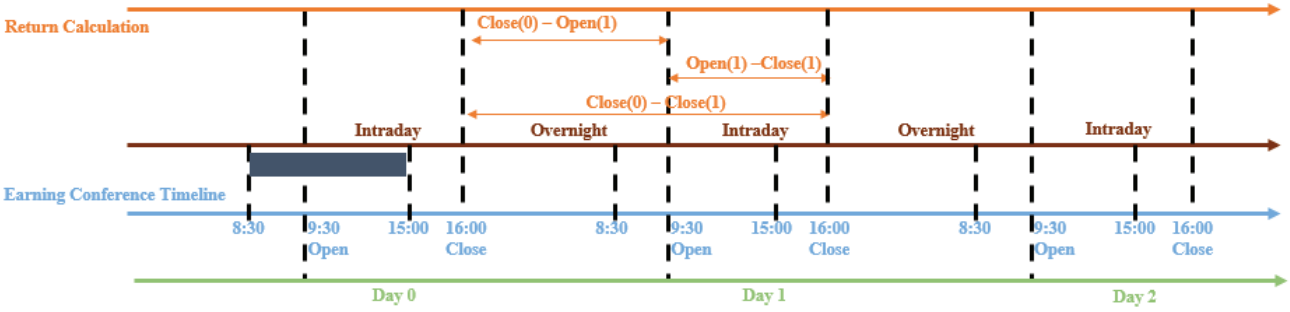

Figure 10: Earning Conference Timeline (during 9:00 to 15:00)

Table 13: Earning Conference during 9:00 to 15:00

	Day 0 close to Day 1 close			Day 0 close to Day 1 open			Day 1 open to Day 1 close		
	L	S	LS	L	S	LS	L	S	LS
Excess Return	0.00%	0.34%	0.23%	0.00%	0.25%	0.31%	-0.12%	0.00%	-0.10%
Standard Deviation	1.40%	2.04%	2.57%	0.51%	1.65%	1.81%	1.36%	1.82%	2.24%
Sharpe Ratio	-0.06	0.16	0.09	0.11	0.15	0.17	-0.09	0.03	-0.04
TSTAT	-0.50	1.44	0.79	0.95	1.33	1.49	-0.80	0.25	-0.38
PVALUE	0.62	0.16	0.43	0.34	0.19	0.14	0.43	0.80	0.70

For earning conferences that take place after the market close (Table 14), strategies that hold the position from open to open underperform the intraday strategies. Both the short-only and long-short portfolios of the intraday strategy gain a Sharpe Ratio that exceeds the infeasible Day 0 strategy. Besides, the portfolios that are built at the close of the next day and held overnight generate negative excess returns, suggesting that the speech sentiment information is fully incorporated during the trading day right after the earning conference is held and the prices exhibit a reversal tendency thereafter.

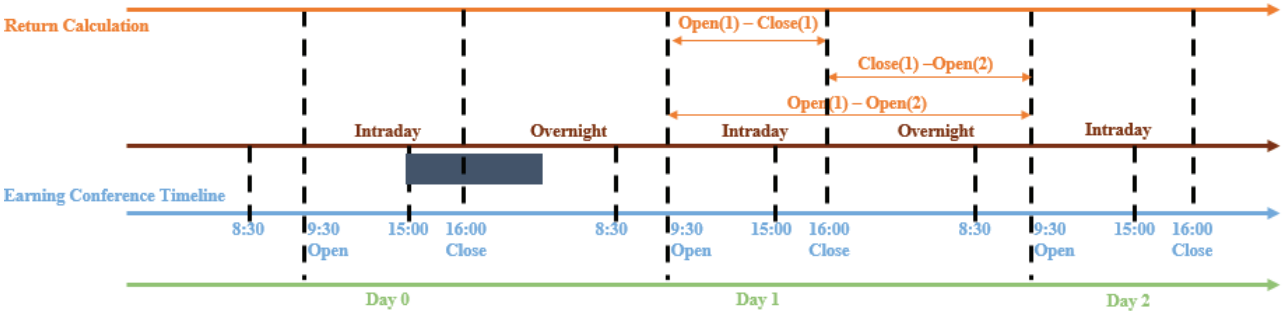

Figure 11: Earning Conference Timeline (after 15:30)

Table 14: Earning Conference after 15:30

	Day 1 open to Day 2 open			Day 1 open to Day 1 close			Day 1 close to Day 2 open		
	L	S	LS	L	S	LS	L	S	LS
Excess Return	0.18%	0.68%	0.83%	0.39%	0.75%	1.11% *	-0.14%	-0.13%	-0.28% **
Standard Deviation	3.43%	3.39%	4.30%	3.58%	3.42%	4.64%	0.72%	0.88%	1.00%
Sharpe Ratio	0.05	0.20	0.19	0.11	0.22	0.24	-0.20	-0.15	-0.28
TSTAT	0.39	1.45	1.41	0.80	1.59	1.74	-1.47	-1.10	-2.04
PVALUE	0.70	0.15	0.17	0.43	0.12	0.09	0.15	0.28	0.05

3.5 Stock Heterogeneity Analysis

In this section, I further investigate differences in sentiment sensitivity and price assimilation with respect to heterogeneity among stocks, in particular the market capitalization.

Previous studies argue that larger firms are less likely to exhibit strong sensitivity to stock price shocks (Chen et al., 2007). Additionally, the speed of price assimilation is likely to be higher for small firms than for large firms. Merton et al. (1987) and Grossman and Miller (1988) show that arbitrage capacity may be less in small-capitalization stocks, and if there are shocks, this could lead to a greater tendency toward reversals.

I form the 9 sentiment-size stock portfolios following Fama and French (1993). Table 15 reports the descriptive statistics. The 9 stock portfolios formed on size and sentiment produce a wide range of average excess returns from -1.13% to 1.64%. In small and medium-sized portfolios, average returns tend to increase with sentiment scores. In large-size portfolios, the return difference between positive and negative sentiment is trivial. Meanwhile, the standard deviations for small-size firms are greater than large-size firms. These risk-return patterns suggest that speech sentiment has a larger impact on small firms than on large firms.

Table 15: Excess returns on 9 stock portfolios formed on sentiment scores and size

	Sentiment Scores quantiles					
Size	Negative	Medium	Positive	Negative	Medium	Positive
quantile	Means			Standard deviations		
Small	-1.13%	-0.59%	1.64%	4.60%	4.25%	5.76%
Medium	-0.72%	0.07%	0.67%	3.50%	2.19%	3.82%
Big	-0.37%	-0.13%	-0.29%	3.62%	1.73%	2.46%
	t-statistics for means			p-values for means		
Small	-1.81	-1.12	2.21	0.08	0.27	0.03
Medium	-1.58	0.25	1.33	0.12	0.81	0.19
Big	-0.76	-0.60	-0.91	0.45	0.55	0.36

In Figure 12 visualize the differences in price adjustment based on firm size. Prices of large stocks respond by 10 basis points on the first day after the earning calls, while the price response of small stocks reaches 280 bp on the first day, almost 28 times larger. The price response of small firms is complete after two days (the day three effect is insignificantly different from zero), while large firms take four days to fully incorporate the speech sentiment. These empirical results are consistent with previous research that small firms are more sensitive and exhibit a greater tendency toward reversals.

3.6 Dimensional Sentiment versus Discrete Sentiment

In previous sections, I employ the one-dimensional emotion theory to classify sentiments according to the valence into two categories: positive ad negative. In this section, I apply the discrete emotion theory

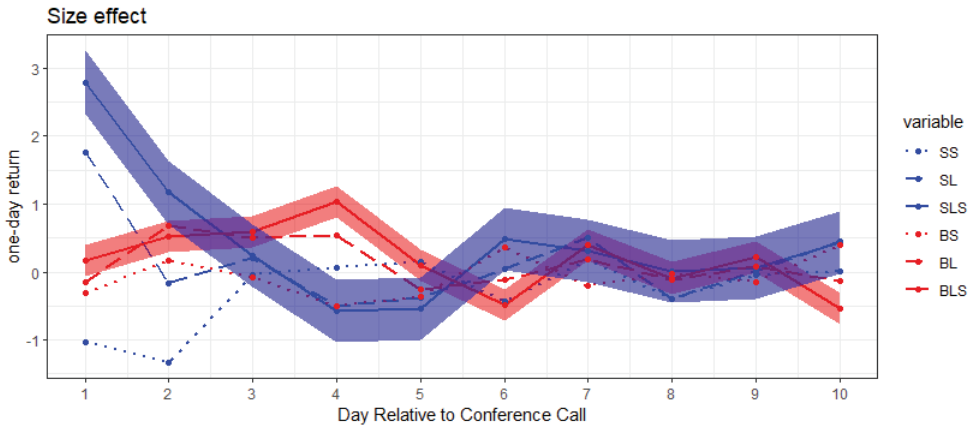

Figure 12: Speed of News Assimilation (Big Versus Small Stocks)

that divides sentiments into six basic categories: sadness, happiness, fear, anger, disgust, and surprise. To train the SETLS, I use the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS Livingstone and Russo 2018) as the source database and extract the discrete sentiment for each utterance with the SETLS model. To aggregate utterance level sentiment into the document level scores, I compute the frequency of each sentiment as the final score for each document. In the 89,843 utterances sample, 71,085 utterances are neutral, 17,821 are happy, 553 are disgust, 257 are fearful, 62 are surprised, 51 are angry and 14 are sad.

Table 16 reports the correlation between the sentiment scores and Day 0/1 open-to-open returns, the corresponding P-value, and the long-short SR for each emotion. It shows that positive emotions (happy and surprised) tend to have positive correlations and Sharpe Ratios on the first day, and negative emotions (sad, disgust, fearful, and angry) tend to have negative or almost zero correlation with the Day 0 return and negative Sharpe Ratios. The top performer 'happy' has a feasible Sharpe Ratio of 0.86, and 'angry' has a feasible Sharpe Ratio of -0.55.

Most of the results are insignificant. This is due partly to the fact that the observations of each sentiment are imbalanced. Approximately 80% of the sample utterances are 'neutral' and merely 0.02 % are classified as 'sad'. While it is natural that most of the time the executives' sentiments during the earning calls are neutral, this highly skewed class proportion of data makes it less reasonable to construct any trading strategy based on discrete emotions. A further potential explanation is that these discrete categories of emotions are not able to define some of the complex emotional states observed in earning calls communication, such as the 'nervous' or 'stressful' led by the deceptive discussions. Though in daily life, this model is intuitive to be used, in this financial scenario, the dimensional emotion model that classifies emotions into positive or negative is more practical. For instance, some of the emotions should be viewed as identical, such as fear and anger, and come emotions like surprise cannot convey effective information to investors since surprise emotion may have positive or negative valence depending on the context.

Table 16: Discrete Sentiment Performance

	Day 0				Day 1			
	Corr		LS		Corr		LS	
	Mean	Pvalue	SR	Pvalue	Mean	Pvalue	SR	Pvalue
Happy	0.31	0.00***	0.78	0.21	0.03	0.53	0.86	0.18
Surprised	0.03	0.53	0.16	0.46	-0.05	0.26	-0.43	0.05**
Sad	0.01	0.85	0.10	0.18	0.00	0.91	0.02	0.78
Disgust	-0.01	0.77	-0.04	0.70	-0.02	0.68	-0.06	0.56
Fearful	0.00	0.97	-0.10	0.44	0.00	0.99	0.01	0.93
Angry	0.01	0.90	-0.16	0.46	-0.02	0.68	-0.55	0.02**

4 Conclusion

This study proposes and analyzes a new framework, SETLS, for the extraction of sentiment information from speech audio records via transfer learning. Compared to the common scoring approach in the finance literature that relies on the usage of the vocal emotion analysis software, this approach delivers customized sentiment scores for individual research applications with the transparency of architecture and the flexibility in feature and source domain selection. For instance, research on deceptive discussion during the financial disclosure could focus on the Teagor Energy Operator features proposed by Zhou et al. (2001) that are designed in particular to measure stress levels and cognitive dissonance. Research using earning conferences in other languages or focusing on particular emotions can apply specific source corpora.

To demonstrate the usefulness of SETLS, I analyze the speech sentiment of 571 earning conferences on the practical problem of portfolio construction. The resulting speech sentiment scores are powerful predictors for price responses. To quantify the impact of speech sentiment on price movement, I construct simple trading strategies and evaluate the performance in various windows following the report day. In order to check if the nonverbal cues contain additional information other than the content, I compare the performance of strategies based on speech and text sentiment and find that the combination of two signals outperforms any of them. I also demonstrate how this framework can be used to investigate the process of price formation and how it can be adjusted to analyze other emotions using different source datasets.

Bibliography

Jeffery S Abarbanell and Victor L Bernard. Tests of analysts' overreaction/underreaction to earnings information as an explanation for anomalous stock price behavior. *The journal of finance*, 47(3):1181–1207, 1992.

Mehmet Berkehan Akçay and Kaya Oğuz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:

56–76, January 2020. ISSN 01676393. doi: 10.1016/j.specom.2019.12.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0167639319302262.

Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

Victor L Bernard and Jacob K Thomas. Post-earnings-announcement drift: delayed price response or risk premium? *Journal of Accounting research*, 27:1–36, 1989.

Claudia Caffi and Richard W. Janney. Toward a pragmatics of emotive communication. *Journal of Pragmatics*, 22(3):325–373, October 1994. ISSN 0378-2166. doi: 10.1016/0378-2166(94)90115-5. URL https://www.sciencedirect.com/science/article/pii/0378216694901155.

Qi Chen, Itay Goldstein, and Wei Jiang. Price informativeness and investment sensitivity to stock price. *The Review of Financial Studies*, 20(3):619–650, 2007.

Jun Deng, Zixing Zhang, Erik Marchi, and Björn Schuller. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *2013 humaine association conference on affective computing and intelligent interaction*, pages 511–516. IEEE, 2013.

Paul Ekman and Harriet Oster. Facial expressions of emotion. *Annual review of psychology*, 30(1):527–554, 1979.

Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, April 2016. ISSN 1949-3045. doi: 10.1109/TAFFC.2015.2457417. Conference Name: IEEE Transactions on Affective Computing.

Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, February 1993. ISSN 0304-405X. doi: 10.1016/0304-405X(93)90023-5. URL https://www.sciencedirect.com/science/article/pii/0304405X93900235.

Robert W. Frick. Communicating emotion: The role of prosodic features. *Psychological Bulletin*, 97(3):412–429, 1985. ISSN 1939-1455. doi: 10.1037/0033-2909.97.3.412. Place: US Publisher: American Psychological Association.

Stefano Giglio, Bryan T. Kelly, and Dacheng Xiu. Factor Models, Machine Learning, and Asset Pricing. SSRN Scholarly Paper ID 3943284, Social Science Research Network, Rochester, NY, October 2021. URL https://papers.ssrn.com/abstract=3943284.

Sanford J Grossman and Merton H Miller. Liquidity and market structure. *the Journal of Finance*, 43(3):617–633, 1988.

John HL Hansen, Sahar E Bou-Ghazale, Ruhi Sarikaya, and Bryan Pellom. Getting started with susas: a speech under simulated and actual stress database. In *Eurospeech*, volume 97, pages 1743–46, 1997.

Jessen L. Hobson, William J. Mayew, and Mohan Venkatachalam. Analyzing Speech to Detect Financial Misreporting. *Journal of Accounting Research*, 50(2):349–392, 2012. ISSN 1475-679X. doi: 10.1111/j.1475-679X.2011.00433.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-679X.2011.00433.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1475-679X.2011.00433.x.

Jingwen Jiang, Bryan T. Kelly, and Dacheng Xiu. (Re-)Imag(in)ing Price Trends. *SSRN Electronic Journal*, 2020. ISSN 1556-5068. doi: 10.2139/ssrn.3756587. URL https://www.ssrn.com/abstract=3756587.

Zheng Tracy Ke, Bryan T. Kelly, and Dacheng Xiu. Predicting Returns with Text Data. SSRN Scholarly Paper ID 3389884, Social Science Research Network, Rochester, NY, September 2020. URL https://papers.ssrn.com/abstract=3389884.

Swarna Kuchibhotla, Hima Deepthi Vankayalapati, RS Vaddi, and Koteswara Rao Anne. A comparative analysis of classifiers in emotion recognition through acoustic features. *International Journal of Speech Technology*, 17(4):401–408, 2014.

Francisco Lacerda. Money Talks: The Power of Voice : A critical review of Mayew and Ventachalam's The Power of Voice: Managerial Affective States and Future Firm Performance. *PERILUS*, pages 1–10, 2012. URL http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-74478. Publisher: Stockholm University.

Jiazheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. MAEC: A Multimodal Aligned Earnings Conference Call Dataset for Financial Risk Prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3063–3070, Virtual Event Ireland, October 2020. ACM. ISBN 978-1-4503-6859-9. doi: 10.1145/3340531.3412879. URL https://dl.acm.org/doi/10.1145/3340531.3412879.

Xi Li, Jidong Tao, Michael T. Johnson, Joseph Soltis, Anne Savage, Kirsten M. Leong, and John D. Newman. Stress and Emotion Classification using Jitter and Shimmer Features. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4, pages IV–1081–IV–1084, April 2007. doi: 10.1109/ICASSP.2007.367261. ISSN: 2379-190X.

Jen-Chun Lin, Chung-Hsien Wu, and Wen-Li Wei. Error Weighted Semi-Coupled Hidden Markov Model for Audio-Visual Emotion Recognition. *IEEE Transactions on Multimedia*, 14(1):142–156, February 2012. ISSN 1941-0077. doi: 10.1109/TMM.2011.2171334. Conference Name: IEEE Transactions on Multimedia.

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

WILLIAM J. Mayew and MOHAN Venkatachalam. The Power of Voice: Managerial Affective States and Future Firm Performance. *The Journal of Finance*, 67(1):1–43, 2012. ISSN 0022-1082. URL https://www.jstor.org/stable/41419670. Publisher: [American Finance Association, Wiley].

Albert Mehrabian et al. *Silent messages*, volume 8. Wadsworth Belmont, CA, 1971.

Robert C Merton et al. A simple model of capital market equilibrium with incomplete information. 1987.

S. McKay Price, Michael J. Seiler, and Jiancheng Shen. Do Investors Infer Vocal Cues from CEOs During Quarterly REIT Conference Calls? *The Journal of Real Estate Finance and Economics*, 54(4):515–557, May 2017. ISSN 0895-5638, 1573-045X. doi: 10.1007/s11146-016-9557-0. URL http://link.springer.com/10.1007/s11146-016-9557-0.

Yu Qin and Yi Yang. What You Say and How You Say It Matters: Predicting Stock Volatility Using Verbal and Vocal Cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 390–401, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1038. URL https://aclanthology.org/P19-1038.

K Sreenivasa Rao, Shashidhar G Koolagudi, and Ramu Reddy Vempada. Emotion recognition from speech using global and local prosodic features. *International journal of speech technology*, 16(2):143–160, 2013.

James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294, 1977.

Klaus R. Scherer. Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99(2):143–165, 1986. ISSN 1939-1455. doi: 10.1037/0033-2909.99.2.143. Place: US Publisher: American Psychological Association.

Klaus R Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1):227–256, April 2003. ISSN 0167-6393. doi: 10.1016/S0167-6393(02)00084-5. URL https://www.sciencedirect.com/science/article/pii/S0167639302000845.

Björn Schuller, Gerhard Rigoll, and Manfred Lang. Hidden markov model-based speech emotion recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 2, pages II–1. Ieee, 2003.

Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wöllmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131, 2010.

Björn W. Schuller. Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99, April 2018. ISSN 0001-0782, 1557-7317. doi: 10.1145/3129340. URL https://dl.acm.org/doi/10.1145/3129340.

Mani Sethuraman, William J. Mayew, and Mohan Venkatachalam. Analyst Conflict of Interest and Earnings Conference Call Informativeness. *SSRN Electronic Journal*, 2018. ISSN 1556-5068. doi: 10.2139/ssrn.3241786. URL https://www.ssrn.com/abstract=3241786.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep Domain Confusion: Maximizing for Domain Invariance. *arXiv:1412.3474 [cs]*, December 2014. URL http://arxiv.org/abs/1412.3474. arXiv: 1412.3474.

Eddie Wong and Sridha Sridharan. Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification. In *Proceedings of 2001 International Symposium on*

Intelligent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No. 01EX489), pages 95–98. IEEE, 2001.

Linyi Yang, Tin Lok James Ng, Barry Smyth, and Riuhai Dong. HTML: Hierarchical Transformer-based Multi-task Learning for Volatility Prediction. In *Proceedings of The Web Conference 2020*, pages 441–451, Taipei Taiwan, April 2020. ACM. ISBN 978-1-4503-7023-3. doi: 10.1145/3366423.3380128. URL https://dl.acm.org/doi/10.1145/3366423.3380128.

Biqiao Zhang, Emily Mower Provost, and Georg Essl. Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences. *IEEE Transactions on Affective Computing*, 10(1):85–99, 2017.

Shiqing Zhang. Emotion recognition in chinese natural speech by combining prosody and voice quality features. In *International Symposium on Neural Networks*, pages 457–464. Springer, 2008.

G. Zhou, J.H.L. Hansen, and J.F. Kaiser. Nonlinear feature based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing*, 9(3):201–216, 2001. ISSN 1063-6676. doi: 10.1109/89.905995.