What You Say and How You Say It Matters: Predicting Financial Risk Using Verbal and Vocal Cues

Yu Qin
School of Information
Renmin University of China
qinyu.gemini@gmail.com

Yi Yang *
HKUST Business School
Hong Kong University of Science and Technology
imyiyang@ust.hk

Abstract

Predicting financial risk is an essential task in financial market. Prior research has shown that textual information in a firm's financial statement can be used to predict its stock's risk level. Nowadays, firm CEOs communicate information not only verbally through press releases and financial reports, but also nonverbally through investor meetings and earnings conference calls. There are anecdotal evidences that CEO's vocal features, such as emotions and voice tones, can reveal the firm's performance. However, how vocal features can be used to predict risk levels, and to what extent, is still unknown. To fill the gap, we obtain earnings call audio recordings and textual transcripts for S&P 500 companies in recent years. We propose a multimodal deep regression model (MDRM) that jointly model CEO's verbal (from text) and vocal (from audio) information in a conference call. Empirical results show that our model that jointly considers verbal and vocal features achieves significant and substantial prediction error reduction. We also discuss several interesting findings and the implications to financial markets. The processed earnings conference calls data (text and audio) are released for readers who are interested in reproducing the results or designing trading strategy.

1 Introduction

Predicting financial risks of publicly traded companies is of great interest to capital market participants. In finance, stock price volatility, which is the standard deviation of a stock's returns over a period of time, is often used as a measure of financial risks. Unlike directly predicting stock prices, it is uncontroversial in the field of economics that one can predict a stock's volatility level using publicly available information (Bernard et al., 2007). Based on this assumption, a burgeoning body of research, both in finance and computational linguistics, has studied predicting stock volatility using various textual sources, including company disclosed reports (Kogan et al., 2009), public news articles (Tetlock, 2007), company earnings call transcripts (Wang and Hua, 2014), and social media (Ding et al., 2015).

Thanks to technological advances, massive amounts of unstructured multimedia data, such as investor conference audio records and CEO public speech videos, have been archived and can be accessed by institutional and individual investors. Everything CEOs (or other executives) say will be closely examined and analyzed by investors. There are anecdotal evidences that CEO's nonverbal features, such as emotions and voice tones, can also be used to reveal firm's performance. For example, it has been reported that hedge fund companies hire ex-CIA agents trained in reading nonverbal cues to assess public statements by managers [1]. While prior research in speech communication has reported that the vocal cues have the power to strengthen or weaken the verbal message, and vocal cues can reflect speaker's affective states or emotion, little research has studied the interplay of verbal cues (language) and nonverbal cues (voice) and their impact on the financial markets.

To fill the gap, we choose a novel multimodal learning setting of company earnings conference call. Earnings conference calls are the periodic conference calls company executives hold with outside investors and analysts to discuss financial results and answer questions raised by analysts. There are three reasons that we choose earnings conference calls as our research setting. First, almost all of the calls are webcast live, and they are later archived on company investor relation (IR) websites or third-party databases. Therefore, both audio and text modalities are available so that we

*Corresponding author.

[1] MarketWatch website. From CIA to BIA: Spotting execs who bend the truth. Accessed: 2019-06-02

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 390–401
Florence, Italy, July 28 - August 2, 2019.

can align vocal cues with verbal cues in multimodal learning, and examine the interplay of both modalities and their impact on the financial markets. Secondly, company earnings announcements are one of biggest stock-moving events. If company reports an earning that does not meet analyst expectation or the CEO fails to address critical questions during the conference call, it often causes significant stock price moves, i.e. high volatility. Lastly, the audio recording and textual transcripts of company earnings conference calls are publicly accessible so interested readers can reproduce the results.

In our work, we propose a stock volatility prediction pipeline using company earnings conference call audio and text data. We construct a unique dataset containing conference call audio and text data of S&P 500 companies in recent years. We then align each sentence in the call transcript with the corresponding audio recording clip. For the multimodal learning, we propose a Multimodal Deep Regression Model (MDRM). The MDRM model utilizes BiLSTM layer to extract context-dependent unimodal features, and subsequently fuses unimodal features together using another layer of BiLSTM to extract multimodal inter-dependencies for the regression task. We empirically demonstrates that MDRM models outperform other benchmark methods significantly and substantially. More importantly, the empirical results confirm that audio modality (vocal cues) help to improve volatility prediction accuracy and may reveal the fact that market participants listen to not only what CEOs say but also how CEOs say it.

Our contributions can be summarized in two folds. First, we are among the first to study the impact of both verbal and vocal features on financial markets, specifically, stock volatility. Secondly, we empirically show that multimodal learning with audio and text can indeed reduce prediction error, compared to previous work that relies on text only. The interesting finding that vocal cues play a role in stock volatility is worth further exploring. In the next section, we briefly provide institutional background on earnings conference call and its impact on financial markets. In Section 3, we outline related work in financial text regression and multimodal learning. We then present our earnings conference call dataset and how data is processed in Section 4. In section 5, we introduce our multimodal learning framework that fuses verbal and vocal features in a deep model. Experiments results are presented in Section 6. Our experiment results show several interesting findings, which we discuss in Section 7. Finally, we conclude this paper in Section 8.

2 Earnings Conference Call and Post Earnings Announcement Drift (PEAD)

Earnings calls are quarterly conference calls company executives hold with outside investors and analysts to discuss firm overall performance. An earnings call consists of two sections: an introduction section and a question-and-answer section. During the introduction section, executives such as CEOs and CFOs read forward-looking statements and provide their information and interpretation of their firms performance during the quarter. During the question-and-answer section, analysts have the opportunity to request managers to clarify information and solicit additional information that the management team does not disclose in the introduction section. The National Investor Relations Institute reports that 92% of companies conduct earnings calls. Institutional and individual investors listen to the earnings call and spot the tones of executives that portend good or bad news for the company.

Company earnings conference call can often result in significant stock price moves. For example, Facebook's stock price dropped over 20% during its nightmare earnings call (second quarter 2018) when the executives said the company expected a revenue growth slowdown in the years ahead. In finance and accounting research, Post Earnings Announcement Drift (PEAD) is a well documented phenomenon that a stock's abnormal returns drift in the direction of an earnings surprise for several weeks following an earnings announcement (Ball and Brown, 1968; Bernard and Thomas, 1989). Moreover, the finance and accounting literature has shown that the stock price moves are largely due to the market reaction to the earnings announcement. The move is most significant during the earnings conference call when the executives start to take analysts questions. In our work, we focus on using executive's verbal and nonverbal cues in conference calls to predict stock price volatility for days following the calls.

3 Related Work

Our work is closely related with the following two lines of research:

financial risk prediction with multimedia data: It is a received wisdom in economics and finance that one can predict a stock's risk using historical information (Bernard et al., 2007). Various work has studied the problem of financial risk prediction using firm financial reports. A pioneer work (Kogan et al., 2009) shows that simple bag-of-words features in firm annual report (Form 10-Ks) combined with historical volatility can simply outperform statistical models that is built upon historical volatility only. Other work (Tsai and Wang, 2014; Nopp and Hanbury, 2015; Rekabsaz et al., 2017; Theil et al., 2018; Wang and Hua, 2014) also proposes different document representation methods to predict stock price volatility. To the best of our knowledge, none of existing NLP research on stock volatility prediction considers the usage of vocal features from audio data, especially the interplay between vocal and verbal features. In finance research, only two studies (Mayew and Venkatachalam, 2012; Hobson et al., 2012) have examined the executive voice in earnings calls. However, they extract CEO's affective state from a blackbox third-party audio processing software, the validity of which has been seriously questioned (Lacerda, 2012).

multimodal learning: Despite our financial domain, our approach is relevant to multimodal learning using text and audio. Recent studies on speech communication have shown that a speaker's acoustic features, such as voice pitch, amplitude, and intensity, are highly correlated with the speaker's emotion (Bachorowski, 1999), deception or trustworthiness(Sporer and Schwandt, 2006; Belin et al., 2017), anxiety (Laukka et al., 2008) and confidence or doubt (Jiang and Pell, 2017).

Recently, multimodal learning has drawn attentions for different applications, such as sentiment analysis (Zadeh et al., 2016b,a; Poria et al., 2017; Luo et al., 2018), image caption generation (You et al., 2016), suicide risk detection (Scherer et al., 2016), crime drama understanding (Frermann et al., 2018) and human trafficking detection (Tong et al., 2017). To the best of our knowledge, this work presents the first multimodal deep learning model using text and audio features for a financial markets application.

4 Earnings Conference Calls Dataset

In this section, we present dataset details.

4.1 Data Acquisition

Conference call transcripts have been extensively studied in prior research. However, there is no existing conference call audio dataset. Therefore, we set up our S&P 500 Earnings Conference Calls dataset by acquiring audio records and text transcripts from the following two sources.

Earnings Call Transcripts. The earnings call transcripts are obtained from the website `Seeking Alpha`[2]. The transcripts are well labeled, including the name of speaker (executives and analysts) and speech content.

Earnings Call Audio. Given each transcript, we download corresponding audio recording from the website `EarningsCast`[3]. The downloaded audio data does not provide any segmentation or labeling for speakers.

4.2 Data Processing

It is too coarse to extract audio features at the conference call transcript level, and it is also too difficult to segment audio recordings at word level. Therefore, we analyze each conference call at sentence level. That is, we want to represent a conference call as a sequence of sentences with corresponding audio clips.

Since conference call normally lasts for about one hour, determining, for each sentence of the transcript, the time interval (in the audio file) containing the spoken text of the sentence is quite challenging. To tackle this challenge, we propose an Iterative Forced Alignment (IFA) algorithm to align each sentence of the transcript with the audio clip containing the spoken text of the sentence. Due to space limit, we present the details of IFA in Appendix. Furthermore, to avoid interference among different speakers, we select only the sentenece made by the most spoken executive (usually the CEO). After the forced alignment step, for each sentence in the conference call transcript, we obtain the sentence text as well as its corresponding audio clip[4].

[2]https://seekingalpha.com/

[3]https://earningscast.com/

[4]It is worth noting that some third-party data provider companies provide human-annotated transcript text and audio recording alignment. In that case, text-audio forced alignment step may not be necessary.

Textual Features We use pre-trained word embeddings and calculate the arithmetic mean of word vector in each sentence as the sentence representation. We choose the embedding `GloVe-300` (Pennington et al., 2014) pre-trained on Wikipedia and Gigaword 5[5]. Therefore, each sentence is represented as a 300-dimension vector.

Audio Features We use Praat (Boersma and Van Heuven, 2001) to extract vocal features, such as pitch, intensity, jitter, HNR(Harmonic to Noise Ratio) and etc, from audio recordings. A total of 27 vocal features are extracted by Praat.

In summary, for each sentence in an earnings conference call, we generate a 300-dimension text vector and a 27-dimension audio vector to represent verbal and vocal features separately.

Data Statistics We build our dataset by acquiring all S&P 500 companies' quarterly earnings conference calls in 2017. We choose S&P 500 constituent firms as the target for volatility prediction for reasons of importance and tractability. Firms in the S&P 500 index encompass roughly three-quarters of the total U.S. market capitalization. A total of 2,243 earnings conference calls are downloaded from `Seeking Alpha` and `EarningsCast`. We discard conference calls which text-audio alignment is not done properly, using the abovementioned data processing method. The final dataset consists of 576 conference calls, with a total number of 88,829 sentences. It can be seen that we discard a large proportion of raw data because the audio-text alignment is very noisy and is prone to errors. We release our processed earnings conference calls dataset[6] (text and audio) for readers who are interested in reproducing the results.

5 Model

We formalize the problem as a supervised machine learning task. The input data is a company's earnings conference call verbal (textual) features and corresponding vocal (audio) features; This is mapped to a numerical variable which is the company's stock price volatility following the conference call.

Prior research (Kogan et al., 2009; Rekabsaz et al., 2017) uses only shallow machine learning model (such as logistic regression) and bag-of-word features to represent financial documents. In other words, the relation and dependencies among the sentences are largely ignored. However, every sentence in a conference call is spoken at a distinct time and in a particular order. Therefore, it is better to treat a conference call as a sequence of sentences. To this end, like other sequence classification problems, we choose to use a recurrent neural network to capture the sentences relation and dependency.

When multimodal verbal and vocal features are available, it is also important to capture the dependency between different modalities, as the vocal cues either affirm or discredit the verbal message. For example, if a CEO says "we are confident about the future product sales" with a voice that is different from the CEO's base vocal cues, such as increased pitch or pauses, we *may* infer that the CEO is not as confident as he claims. In fact, existing research (Jiang and Pell, 2017) in speech communication has shown that voice (vocal cues) plays a critical role in verbal communication. If we ignore the voice patterns that are accompanied with the verbal language, we may misinterpret the CEO's statement. Especially in financial markets where CEO's word and voice are closely examined by professional analysts and investors, it is plausible that market reacts to both verbal and vocal signals.

Therefore, we present a deep model to capture context-dependent unimodal features and fuse multimodal features for the regression task. The high-level idea behind the design is to use contextual BiLSTM to extract context-dependent unimodal features separately corresponding to each sentence, and then use a BiLSTM to fuse multimodalities and extract the inter-dependencies between different modalities. The details of our model is described below.

5.1 Notations

We first introduce our notations. Let M be the total number of conference call transcripts while the longest one has N sentences. Then we denote $\boldsymbol{X}_j$ as the j_{th} conference call, where $1 \leq j \leq M$. In our multimodal setting, $\boldsymbol{X}_j = [\boldsymbol{T}_j; \boldsymbol{A}_j]$. $\boldsymbol{T_j}$ is a $N \times dt$ matrix that represents the document embeddings of the call transcripts, where N is the number of sentences in a document[7] and dt is the

[5] https://nlp.stanford.edu/projects/glove/

[6] Our dataset is available at `https://github.com/GeminiLn/EarningsCall_Dataset`

[7] Assuming the longest document has N sentences, for documents which contain less than N sentences, we utilize zero-padding to fill them to N to keep consistency.

dimensions of word embedding. $\boldsymbol{A_j}$ is a $N \times da$ matrix that represents the vocal features extracted from earnings call audios, where da is the dimensions audio feature. y_j and $\hat{y_j}$ represent the true and predicted stock volatility value corresponding to j_{th} conference call.

5.2 Multimodal Deep Regression Model

Our multimodal deep regression model (MDRM) includes two components. The first component is a contextual BiLSTM that extracts unimodal features for either text or audio modality. The contextual BiLSTM is able to capture the relationship and dependency for unimodal inputs. In the second component, the extracted multimodal (text and audio) features are then combined and are fed into a BiLSTM with a fully-connected layer, which extracts inter-dependencies between text and audio modality.

5.2.1 Extracting Unimodal Features with Contextual BiLSTM

The Contextual LSTM is proposed by (Poria et al., 2017), designed to analyze video emotion utilizing text, speech and video image. The contextual LSTM connects dense layers and softmax output with each LSTM unit. In the implementation, this architecture is also called time-distributed dense layer. This structure helps maintain the latent time sequence in data while making sentiment classification on the utterance level.

In our contextual LSTM, we choose the BiLSTM as fundamental LSTM architecture by its best performence in past work (Poria et al., 2017). BiLSTM is the bidirectional LSTM (Hochreiter and Schmidhuber, 1997), which is an extended model of recurrent neural network (RNN). Specifically, LSTM is designed to acquire key information from time series data while overcoming the defect that traditional RNN might lose information in long time series. BiLSTM is then developed from LSTM, considering not only the forward information transfer but backward transfer. The bidirectional information transmission significantly improves model prediction power. For the construction of Contextual BiLSTM, detailed formulas (Only forward transmission formulas) are described below.

$$f_j = \sigma_g(W_f x_j + U_f h_{j-1} + b_f)$$

$$i_j = \sigma_g(W_i x_j + U_i h_{j-1} + b_i)$$

$$o_j = \sigma_g(W_o x_j + U_o h_{j-1} + b_o)$$

$$c_j = f_j \circ c_{j-1} + i_j \circ \sigma_c(W_c x_j + U_c h_{j-1} + b_c)$$

$$h_j = o_j \circ \sigma_h(c_j)$$

$$Z_j = ReLU(W_z h_j + b_z)$$

In the above formulas, x_j denotes the j_{th} input features, i.e., the j_{th} sentence textual or audio features. f_j, i_j, and o_j represent the standard forget gate, input gate and output gate. W and b are trainable vectors in the training process, and all the vectors described above are used to generate hidden state h_j and cell state c_j. Z_j in the last formula stands for the output of time-distributed dense layer connected to the j_{th} LSTM unit.

Compared with Poria's work (Poria et al., 2017), we remove the softmax output on LSTM unit since our regression is applied on document level, instead of utterance level. The dense layer output is constructed as a new time sequence feature to be further utilized in next stage.

5.2.2 Hierarchical Fusion of Unimodal Features

Hierarchical fusion of unimodal features is achieved by our Multimodal Deep Regression Model. Figure 1 demonstrates the integral process. In this process, the hierarchical fusion consists of two stages.

Stage 1 Vectors $\boldsymbol{T}$ and $\boldsymbol{A}$ are represented by the matrices on the left. Matrix $\boldsymbol{T}$ is 520×300 dimensional and matrix $\boldsymbol{A}$ is 520×27 dimensional, while 520 is the length of document, 300 and 27 are the dimensions of textual features and audio features. The matrices are then fed into Contextual BiLSTM through a Mask layer to screen the effect of zero-padding. As described in 5.2.1, Contextual BiLSTM extracts unimodal features for each matrix separately while keep the original chronological order. After extracted, unimodal features are still organized on sentence level so they can be horizontally stitched as merged features in the middle of Figure 1.

Stage 2 The merged features are then fed into a BiLSTM connected with a two-layer neural network. To be specifically, we avoid the same network architecture as Poria's work (Poria et al., 2017) here to achieve our unique purpose. Unlike video emotion classification, the regression problem in our study is document-level, which means that we do not make prediction on each utterance. Therefore, Contextual BiLSTM is not suitable for

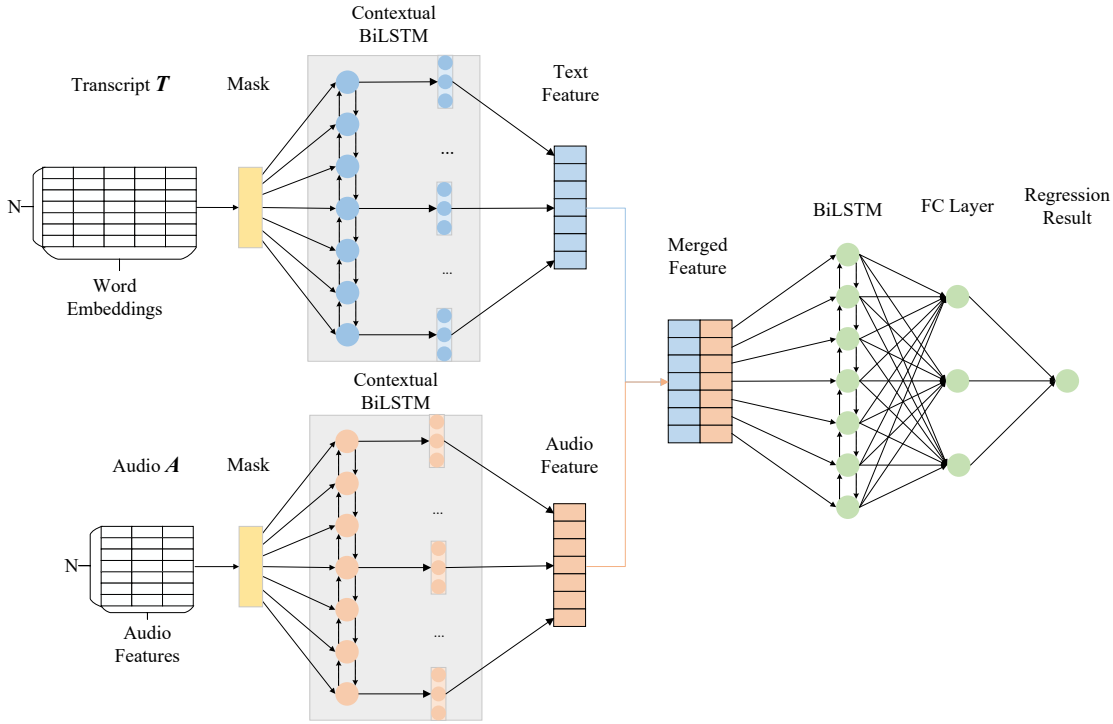

Figure 1: The proposed Multimodal Deep Regression Model (MDRM). The inputs to the model is a company's conference call audio file with correpsonding transcript. Each conference call consists of N sentences. The output variable is a numerical value, i.e., the company's stock price volatility following the conference call.

stage 2 since the features are already extracted on high-level. In stage 2, we use the BiLSTM connected with a two-layer neural network to complete the regression. The effectiveness of this concision structure will be experimental proved in the experiment result section.

6 Experiment Setup

The stock volatility prediction problem is formulated following (Kogan et al., 2009). The volatility is defined as:

$$v_{[t-\tau,t]} = \ln\left(\sqrt{\frac{\sum_{i=0}^{\tau}(r_{t-i}-\bar{r})^2}{\tau}}\right) \quad (1)$$

where r_t is the return price at day t and $\bar{r}$ is the mean of the return price over the period of day $t-\tau$ to day t. The return price is defined as $r_t = \frac{P_t}{P_{t-1}} - 1$, where P_t is the closing price on day t. We choose different τ values, including 3, 7, 15, 30 calendar days to evaluate the short-term and long-term effectiveness of volatility prediction. We obtain daily stock prices of year 2017 (dividend-adjusted) from CRSP database.

We report the performance using the Mean Squared Error (MSE) between the predicted volatility and true volatility:

$$MSE = \frac{1}{M'}\sum_{i=1}^{M'}(f(\boldsymbol{X}'_i) - y'_i)^2 \quad (2)$$

where M' is the size of the test set, and y'_i is the true volatility associated with testing example $\boldsymbol{X}'_i$.

6.1 Baselines

We consider several stock volatility prediction baselines as described below.

Past Volatility. It is often reported in prior research that past volatility is a strong predictor of future volatility. Thus we consider using the volatility of previous τ-days before conference call to predict the τ-days volatility following the conference call. We call this baseline v^{past}.

tf-idf bag-of-words. It is used in (Kogan et al., 2009). The feature value is classic `tf-idf` score. Term frequency (tf) is calculated as $TF = \frac{n_{i,j}}{\sum_k n_{k,j}}$, and inverse document frequency (idf) is calculated as $IDF = \log\left(\frac{|d|}{1+df(t)}\right)$, where the $n_{i,j}$ is the number of frequency of term t_i in document d_j, and $\sum_k n_{k,j}$ denotes the sum of all terms appear in document d_j. $|d|$ is the total number of document, and $df(t)$ is the sum of documents which contain the term t_i.

word embeddings. Each transcript is represented as a weighted average of word embeddings. In our experiment, we use pre-trained GloVe-300 word embeddings. This document representation is shown to be a simple yet effective method (Arora et al., 2017). This baseline can help us to evaluate the effectiveness of proposed deep model. We also experiment with pre-trained word embeddings GloVe-50 and GloVe-100 but find GloVe-300 performs the best among those. Therefore, we use GloVe-300 as input word embeddings throughout our experiments.

For the above two baselines tf-idf bag-of-words and word embeddings, given conference call transcript representations, we apply Support Vector Regression (SVR) (Drucker et al., 1997) with Radial Basis Function (RBF) kernel to predict stock volatility y_i, following previous studies (Kogan et al., 2009; Rekabsaz et al., 2017; Tsai and Wang, 2014).

We also consider two multimodal learning baselines that fuse both audio and textual features.

simple fusion This is a baseline using a simple shallow model to fuse different modalities. The audio and text features are fed into SVR as input. Using this baseline, we can compare the effectiveness of deep multimodal model with shallow multimodal model.

bc-LSTM It is a state-of-the-art multimodal learning model as proposed in (Poria et al., 2017). They present a bidirectional contextual LSTM (**bc-LSTM**) framework for fusing multimodal features including audio, video and text. We replicate their deep model as a direct baseline.

For our multimodal deep regression model (MDRM), we also evaluate three different scenarios: **text-only**, **audio-only**, and both text and audio are available **text+audio**.

6.2 Training Setup

Our deep model is built and trained with Keras[8]. We apply backpropagation with stochastic gradient descent in the training, and we choose the mean square error as the loss function. We use linear activation for the final regression layer and implement ReLU activation function for the remaining layers.

During the experiment, we find that training with audio data is more prone to overfitting. We then implement dropout in our model. In the first stage, we set dropout as 0.5 for audio contextual BiLSTM and 0.8 for text contextual BiLSTM. In the second stage, we remove the dropout layer. For the model evaluation, randomly splitting dataset into training/validation/testing is not reasonable since we should not use later years' conference calls to predict previous years' stock volatilities. Therefore, we choose the top 80% of the data as training data and the remaining 20% as test data.

[8]Keras: https://keras.io/

7 Experiment Results and Discussion

Predicting stock volatility is a rather challenging task given the noisiness of the stock markets. Following prior research, we report volatility number in the 3-th decimal. The main experiment results are shown in Table 1. We now discuss the experiment results and several interesting findings as well as their implications to the stock markets.

Multimodal Deep Regression Model is Effective. The results show that our multimodal deep regression model (MDRM) outperforms all baselines. Using both text and audio data, the model has prediction error of 1.371, 0.420, 0.300 and 0.217 for 3-days, 7-days, 15-days and 30-days following the conference call respectively. Comparing with using past volatility only, the improvement gain is as substantial as 54.1% for 3-days prediction. The improvement over other baseline methods are 19.1% (tf-idf bag-of-words), 17.8% (word embeddings), 20.4%(simple fusion) respectively for 3-days prediction. Comparing with the state-of-art baseline bc-LSTM (Poria et al., 2017), MDRM also achieve 3.3% error reduction for 3-days prediction. It is worth emphasizing the substantial improvement over simple fusion model. As our design motivation, verbal and vocal features should be modeled jointly as vocal cues either affirm or discredit the verbal message in public communication. Our deep regression model is able to capture the interplay of both modalities that a simple feature fusion model cannot.

Both modalities are helpful. We can also conclude from the results that multimodal features are more helpful than unimodal features (either text or audio) alone. When we predict the stock volatility 3-days following the conference call, multimodal (1.371) outperform unimodal (1.431) by 4.2%. As shown in Table 1, MDRM (text+audio) significantly outperforms MDRM (text only) and MDRM (audio-only) model for 3-days, 7-days and 15 days stock volatility prediction. The im-

		τ=3	τ=7	τ=15	τ=30
v^{past}		2.986	0.826	0.420	0.231
tf-idf bag-of-words		1.695	0.498	0.342	0.249
word embeddings		1.667	0.549	0.345	0.275
simple fusion		1.722	0.501	0.307	0.233
bc-LSTM (text+audio) (Poria et al., 2017)		1.418	0.436	0.304	0.219
Multimodal Deep Regression Model (MDRM)	text only	1.431	0.439	0.309	0.219
	audio only	1.412	0.440	0.315	0.224
	text+audio	1.371***	0.420***	0.300**	0.217

Table 1: MSE of different models on stock volatility prediction τ-days following the conference call. The * denotes statistical significance compared to MDRM (text only) results under a one-tailed t-test (*** for $p \leq 0.001$ and ** for $p \leq 0.01$)

provement is not statistically significant for 30-days prediction, which we will explain the possible reasons later. In addition to reduced prediction error, fusing both modalities can mitigate potential overfitting problem. We find that training a deep LSTM network with audio data only can result in overfitting very quickly. In our experiment, the audio-only deep network shows a trend of over-fitting in 10 epochs. Therefore, the result that audio-only MDRM performs better than text-only MDRM (1.412 vs. 1.431) may need careful interpretation as we have to stop audio-only model training early to prevent overfitting. However, using both audio features and text features, the model usually converges in 20 epochs without over-fitting.

Some Individual Vocal Features are Important. We also design another experiment to investigate the importance of different vocal features. We examine whether the left-out of individual vocal features can affect prediction results. We follow the prior research (Jiang and Pell, 2017) to select five representative vocal features including mean pitch, standard deviation of pitch, mean intensity, number of pulses and mean HNR (Harmonic-to-Noise Ratio). Our experiment results show that without mean pitch feature, the MSE of our model increases 0.7%. The left-out of standard deviation of pitch also raises MSE by 0.65%. For mean intensity and number of pulses, MSE increases by 0.63% and 0.56% respectively. However, MSE is not changed with mean HNR being left-out.This finding is consistent with prior research in speech communication that pitch and intensity are important features when detecting a speaker's confident and doubt.

Short-term Volatility Prediction is Hard. Our prediction results consistently show that short term volatility prediction error is much greater than long term prediction error. For example, the 3-days prediction MSE of MDRM is 1.371, while the 30-days MSE is 0.217. The gain of MDRM over past volatility baseline v^{past} diminishes from 54% ($\tau = 3$) to 6% ($\tau = 30$). In other words, short term volatility prediction is much more difficult than long term prediction. This phenomenon has also been extensively documented in finance and accounting literature, known as post earnings announcement drift (PEAD). Research (Ball and Brown, 1968; Bernard and Thomas, 1989) have shown that the stock price moves more significantly (volatile) in a short period of time (several trading days) following the conference call than in a long period of time (from weeks to months). Even though the absolute value of MSE is higher in short-term, the 54% improvement over baseline past volatility is still encouraging, because any information that helps to formulate realistic estimates of the volatility can be invaluable to capital market participants.

Marginal Gain over Simple Models is Diminishing in Long-term. Our experiment results also consistently show that complex deep models such as bc-LSTM (Poria et al., 2017) or our proposed deep regression model outperform shallow models (such as SVR) by large margin in short-term prediction (τ=3 or 7). However, the margin becomes smaller as we predict a relative long-term stock volatility (τ=15 or 30). For example, comparing with tf-idf bag-of-words model at $\tau = 3$, our MDRM reduces prediction error by 19.1% (1.371 vs. 1.695). However, at $\tau = 30$, the prediction error reduction is 12.8% (0.217 vs. 0.249). This can also be confirmed that when $\tau = 30$,

the MSE of past volatility method is as small as 0.231, which is even better than tf-idf bag-of-words model and is only slightly worse than MDRM. In other words, the benefit of using complex deep model for long-term volatility prediction is smaller than for short-term volatility prediction. This phenomenon can be explained by Efficient-market hypothesis (EMH), which is a theory in financial economics that states that the stock prices only react to new information so it is impossible to predict the stock price based on historical information. Therefore, as we target for a longer time horizon, the predictive power of using the previous conference calls information becomes less significant and substantial.

7.1 Case Study: AMD Conference Call First Quarter 2017

We conduct a case study to further investigate the validity of multimodal learning for stock volatility prediction. The case study is based on the AMD (Advanced Micro Devices Inc.)'s earnings conference call in the first quarter of 2017. We qualitatively explain why multimodal features are more helpful than unimodal text features.

May 1st 2017 is a bad day for AMD investors. After the company's earnings conference call, the stock price dropped by 16.1% in the post market session. The company's stock price became very volatile for the next few days. We analyze the conference call transcript with corresponding audio recording of the company's Chief Executive Officer (CEO) Dr. Lisa T. Su.

Figure 2 illustrates the inconsistencies between the CEO's verbal cues and her vocal cues. We observe that there is a significant increase in mean pitch while the CEO is saying "*Overall, from a performance standpoint, the product and the customer engagements are going as we would expect*" (Case 1). While the language is positive, the mean pitch of CEO's voice increases 20% above her average mean pitch (203.39 Hz) and the mean pitch values in nearby sentences. According to prior acoustic research (Jiang and Pell, 2017), the high mean pitch may correlate with a speaker being not confident about what he or she is talking about. A similar inconsistency also happens when the CEO is saying *We have more memory bandwidth*" (Case 2).

After the earnings conference call, it turns out that the revenue of AMD actually missed the analyst expectation by $0.38M. Thus, the positive words in the CEO's language is not as credible as it sounds. Using unimodal text data only, we may miss the inconsistency in verbal and vocal cues. Therefore, the multimodal learning model may capture the inter-dependency between multimodal features and better predict market reactions to earnings conference calls.

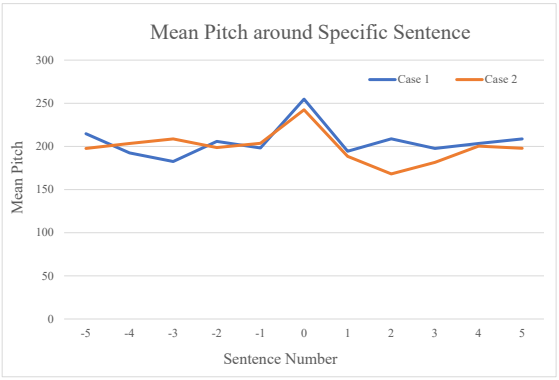

Figure 2: The change of Mean Pitch around specific sentence. Sentence with number 0 is the corresponding Case1 and Case2 sentence described in the paper.

8 Conclusion

Predicting financial risks of publicly traded companies is an essential task in financial markets. In this work, we have demonstrated that CEO's language and voice in company earnings conference calls can be utilized to predict the company financial risk level, as measured by stock price volatility for days following the conference call. We propose a BiLSTM-based multimodal deep regression model that extracts and fuses multimodal features from text transcripts and audio recordings. Even though our work is an application of financial domain, we hope our multimodal learning model can also be useful in other areas (such as social media and customer service) where multimodality data is available.

Acknowledgments

This work was supported by Theme-based Research Scheme (No. T31-604/18-N) from Research Grants Council in Hong Kong, and the National Natural Science Foundation of China (Grant No. 71771212, U1711262). We thank the anonymous reviewers for helpful comments. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *In Proceedings of ICLR*.

Jo-Anne Bachorowski. 1999. Vocal expression and perception of emotion. *Current directions in psychological science*, 8(2):53–57.

Ray Ball and Philip Brown. 1968. An empirical evaluation of accounting income numbers. *Journal of accounting research*, pages 159–178.

Pascal Belin, Bibi Boehme, and Phil McAleer. 2017. The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. *PloS one*, 12(10):e0185651.

Dumas Bernard, Kurshev Alexander, and Uppal Raman. 2007. Equilibrium portfolio strategies in the presence of sentiment risk and excess volatility. Working Paper 13401, National Bureau of Economic Research.

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

Paul Boersma and Vincent Van Heuven. 2001. Speak and unspeak with praat. *Glot International*, 5(9/10):341–347.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *In Proceedings of IJCAI*, pages 2327–2333.

Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex J. Smola, and Vladimir Vapnik. 1997. Support vector regression machines. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *In Proceedings of NIPS*, pages 155–161.

Lea Frermann, Shay B Cohen, and Mirella Lapata. 2018. Whodunnit? crime drama as a case for natural language understanding. *Transactions of the Association of Computational Linguistics*, 6:1–15.

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

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Xiaoming Jiang and Marc D Pell. 2017. The sound of confidence and doubt. *Speech Communication*, 88:106–126.

Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. Predicting risk from financial reports with regression. In *In Proceedings of NAACL*, pages 272–280.

Francisco Lacerda. 2012. Money talks: The power of voice: A critical review of mayew and ventachalams the power of voice: Managerial affective states and future firm performance. *PERILUS*, pages 1–10.

Petri Laukka, Clas Linnman, Fredrik Åhs, Anna Pissiota, Örjan Frans, Vanda Faria, Åsa Michelgård, Lieuwe Appel, Mats Fredrikson, and Tomas Furmark. 2008. In a nervous voice: Acoustic analysis and perception of anxiety in social phobics speech. *Journal of Nonverbal Behavior*, 32(4):195.

Ziqian Luo, Hua Xu, and Feiyang Chen. 2018. Utterance-based audio sentiment analysis learned by a parallel combination of cnn and lstm. *arXiv preprint arXiv:1811.08065*.

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

Pedro J Moreno, Chris Joerg, Jean-Manuel Van Thong, and Oren Glickman. 1998. A recursive algorithm for the forced alignment of very long audio segments. In *In proceedings of ICSLP*.

Clemens Nopp and Allan Hanbury. 2015. Detecting risks in the banking system by sentiment analysis. In *In Proceedings of EMNLP*, pages 591–600, Lisbon, Portugal.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *In Proceedings of EMNLP*, pages 1532–1543.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *In Proceedings of ACL*, volume 1, pages 873–883.

Navid Rekabsaz, Mihai Lupu, Artem Baklanov, Allan Hanbury, Alexander Duer, and Linda Anderson. 2017. Volatility prediction using financial disclosures sentiments with word embedding-based ir models. *In Proceedings of ACL*, pages 1712–1721.

Stefan Scherer, Gale M Lucas, Jonathan Gratch, Albert Skip Rizzo, and Louis-Philippe Morency. 2016. Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews. *IEEE Transactions on Affective Computing*, 7(1):59–73.

Siegfried Ludwig Sporer and Barbara Schwandt. 2006. Paraverbal indicators of deception: A meta-analytic synthesis. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 20(4):421–446.

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

Christoph Kilian Theil, Sanja Stajner, and Heiner Stuckenschmidt. 2018. Word embeddings-based uncertainty detection in financial disclosures. In *In Proceedings of the First Workshop on Economics and Natural Language Processing*, pages 32–37.

Edmund Tong, Amir Zadeh, Cara Jones, and Louis-Philippe Morency. 2017. Combating human trafficking with multimodal deep models. In *In Proceedings of ACL*, pages 1547–1556.

Ming-Feng Tsai and Chuan-Ju Wang. 2014. Financial keyword expansion via continuous word vector representations. In *In Proceedings of EMNLP*, pages 1453–1458.

William Yang Wang and Zhenhao Hua. 2014. A semiparametric gaussian copula regression model for predicting financial risks from earnings calls. In *In Proceedings of ACL*, volume 1, pages 1155–1165.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016a. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016b. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

A Appendices

In this appendix section, we present details of our text and audio forced alignment method. Given an audio file containing speech, and the corresponding transcript, forced alignment is defined as the process of determining, for each fragment of the transcript, the time interval (in the audio file) containing the spoken text. In our setting, we need to match speaker's speech and corresponding spoken text from an earnings conference call data.

However, earnings conference call normally lasts for about one hour or longer. Therefore, aligning audio clips with the corresponding text is quite challenging.

Toward this end, we propose an Iterative Forced Alignment (IFA) algorithm to promote the alignment results on our data set. The IFA method is inspired by a spoken language processing work (Moreno et al., 1998). We implement IFA on the basis of normal forced alignment technology, in Python, we use Aeneas[9] as fundamental forced alignment method. Algorithm 1 demonstrates the specific architecture of our method.

Algorithm 1 Iterative Forced Alignment

function $Alignment(a_i, t_i, s_i)$
  if $Length(a_i) = 0$ **then**
    return $True$
  end if
  if $Length(a_i)! = 0$ **then**
    $result \leftarrow Aeneas(a_i, t_i)$
    $speaker \leftarrow LastSpeaker(s_i)$
    $slice_{a,t} \leftarrow LastParagraph(a_i, t_i)$
    $s_i \leftarrow CutLastSpeaker(s_i)$
    $a_i, t_i \leftarrow CutLastParagraph(a_i, t_i)$
    Save $slice_{a,t}$ as files
    return $False$
  end if
end function
function $IterativeSegmentation$
  for $i = 0 \rightarrow M$ **do** ▷ M is the number of calls
    $a_i, t_i \leftarrow Audio_i, Trasncript_i$
    $s_i \leftarrow SpeechSequence_i$
    while $result! = True$ **do**
      $result \leftarrow Alighment(a_i, t_i, s_i)$
    end while
  end for
end function

During our experiment, we find the forced alignment performs well in the beginning and end of the whole document. In the middle parts, alignment result might be influenced by short syllable words, fast switching of speakers or omission of text record. Therefore, we utilize the iterative strategy in segmentation. Instead of aligning the whole document and then segment it according to alignment result, the IFA chooses to segment only the last paragraph at one time, since the last paragraph is most likely to be aligned precisely. After segment the last paragraph, IFA will restart the forced alignment on the remaining audio and text, generate the new alignment result and segment the last paragraph, until document is fully processed. We randomly select 200 earnings conference calls to test the effectiveness of IFA. As shown in Table 2, the adoption of IFA improves segmentation accuracy and reduces the degree of error significantly.

[9] Aeneas: https://github.com/readbeyond/aeneas

	Match		Not Match	
	Begin	End	Begin	End
Iterative	63	60	37	40
	Total:123		Total:77	
One-Time	33	22	67	78
	Total:55		**Total:145**	

Table 2: Comparison of Iterative Segmentation and One-Time Segmentation

To acquire right-segmented earnings conference calls automatically. We implement both IFA and One-Time segmentation on the remaining data, selecting the right-segmented earnings conference call by comparing the result of two methods. If the difference of segmentation result between the two methods is small in one document, we note this document as right-segmented.

By adopting IFA on our dataset, we solve the long, noisy audio segmentation problem in an effective way. Since there is no recognized practical method to deal with such a problem, our work can contribute to those researchers who are interested in long audio processing and analyzing. Not only in financial materials analysis field but also in other areas including social media analysis and emotion recognition.