VolTAGE: Volatility Forecasting via Text Audio Fusion with Graph Convolution Networks for Earnings Calls

Ramit Sawhney
Netaji Subhas Institute of Technology
ramits.co@nsit.net.in

Piyush Khanna
Delhi Technological University
piyushkhanna_bt2k17@dtu.ac.in

Arshiya Aggarwal
MIDAS, IIIT Delhi
arshiya.dtu@gmail.com

Taru Jain
MIDAS, IIIT Delhi
jaintaru@ieee.org

Puneet Mathur
University of Maryland, College Park
puneetm@cs.umd.edu

Rajiv Ratn Shah
IIIT Delhi
rajivratn@iiitd.ac.in

Abstract

Natural language processing has recently made stock movement forecasting and volatility forecasting advances, leading to improved financial forecasting. Transcripts of companies' earnings calls are well studied for risk modeling, offering unique investment insight into stock performance. However, vocal cues in the speech of company executives present an underexplored rich source of natural language data for estimating financial risk. Additionally, most existing approaches ignore the correlations between stocks. Building on existing work, we introduce a neural model for stock volatility prediction that accounts for stock interdependence via graph convolutions while fusing verbal, vocal, and financial features in a semi-supervised multi-task risk forecasting formulation. Our proposed model, VolTAGE, outperforms existing methods demonstrating the effectiveness of multimodal learning for volatility prediction.

1 Introduction

Motivation Financial risk modeling is of great interest to capital market participants for making sound investment decisions. Stock volatility is a vital indicator of a company's risk profile (Poon and Granger, 2003; Yang et al., 2020). The stock market presents various opportunities that increasingly attract investors, who utilize the market's potential to generate profits, wherein stock volatility is a vital risk modeling factor. One underexplored, yet crucial event that leads to significant fluctuations in stock volatility, is the earnings conference call. These calls are held periodically by publicly traded companies' executives to summarize and prognosticate company's performance (Qin and Yang, 2019). Harnessing the interplay between the multimodal verbal and vocal cues in earnings calls can help better analyze the impact these calls may have on financial markets and forecast stock volatility (Dichev and Tang, 2009; Yang et al., 2020).

Challenges While stock trading presents unparalleled investment opportunities, accurately predicting the rise and fall of stock prices has numerous challenges (Campbell et al., 1997). Conventional research in finance revolves around using historical stock data to develop statistical models and recurrent neural networks (RNNs) capable of forecasting price trends (Kristjanpoller et al., 2014; Zheng et al., 2019). They are influenced by many factors ranging from public opinion to the movements of other related stocks (Malkiel, 2003). Recent advances in deep learning present a promising prospect in multimodal stock forecasting by analyzing online news (Hu et al., 2018), and social media (Guo et al., 2018) to learn latent patterns affecting stock prices (Jiang, 2020). However, the challenging aspect in stock forecasting is that most existing work treats stock movements to be independent of each other, contrary to true market function (Diebold and Yılmaz, 2014). Additionally, existing research has not leveraged the rich audio signals in company executives' speech, which could indicate the emotional and affective state of the speakers, and provide insights into company performance. More recently, the use of audio processing for earnings calls has gained an interest in both financial and linguistic research (Burgoon et al., 2015; Jiang and Pell, 2017).

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 8001–8013,
November 16–20, 2020.

Multimodal approaches can extract complementary information from multiple modalities to improve financial modeling, MDRM (Qin and Yang, 2019), and HTML (Yang et al., 2020) validate the premise of such approaches for volatility forecasting. Additionally, advances in graph-based deep learning (Kipf and Welling, 2017) have led to the rise of graph neural networks (GNNs) that can model the relationships between related stocks (Feng et al., 2019). Publicly available online company information can be used to identify connections between stocks that might influence each other, such as those having the same CEO or belonging to the same industry. Financial tasks are often correlated, thus making multi-task learning a promising choice for financial forecasting.

Contributions Building on advances in the intersection of financial research, graph neural networks, and natural language processing, we present VolTAGE: **Vol**atility forecasting via **T**ext-**A**udio fusion with **G**raph convolution networks for **E**arnings calls. VolTAGE comprises a set of neural components to capture cross-modal signals from earnings calls transcripts, CEO speech, inter stock dependence graphs, and numerical financial features. First, VolTAGE combines the verbal-vocal coherence between earnings calls transcripts and speech via an inter-modal multi-utterance attention mechanism. The fused features are then fed to a graph convolution network (GCN) to simultaneously solve two homogeneous stock volatility tasks - average volatility (main task) and single-day volatility prediction (auxiliary task), in a semi-supervised fashion. Through a set of comparative, qualitative, and ablation experiments on real-world S&P 500 index data, we show VolTAGE's utility of augmenting vocal and verbal cues with graph-based features in a multi-task setup.

Ethical Considerations and Limitations Examining a CEO's speech and tone in earnings calls is a well-studied phenomenon in financial literature (Crawford Camiciottoli, 2011; Qin and Yang, 2019). Our work focuses only on calls for which companies publicly release transcripts and audio recordings. The data used in our study corresponds to earnings calls of S&P 500 companies. We acknowledge the presence of gender bias in our study, given the imbalance in the gender ratio of CEOs of S&P 500 companies. We also acknowledge the demographic bias in our study, as the S&P 500 companies are organizations listed in the US, and may not generalize directly to non-native speakers.

2 Background

Extensive studies have shown the utility of employing historical financial data (Jones, 2017; Dichev and Tang, 2009) for volatility prediction, yet financial forecasting using multiple modalities remains an underexplored avenue. While newer work focuses on data across multiple modalities, there exist drawbacks and understudied approaches to improve current methods, which we describe next.

Volatility Forecasting Forecasting stock volatility is a crucial pillar across multiple financial domains and has focused on numerous academic studies. Volatility is a key indicator of uncertainty and is a decisive variable to many investment decisions and portfolio creations. Previous work in this domain has mainly relied on numerical features (Liu and Chen, 2019; Nikou et al., 2019), such as macroeconomic indicators (Hoseinzade et al., 2019). This includes discrete (GARCH (Duan, 1995), rolling regression (Peng et al., 2018)), continuous (Andersen, 2007), and neural approaches (Kogan et al., 2009). This comprehensive work illustrates the significance of volatility in investment, security valuation, and risk management.

Natural Language Processing and Finance Extensive studies incorporating related text information have proven successful in financial forecasting tasks. Mohan et al. (2019); Tan et al. (2019) utilized financial news articles to improve the accuracy of stock price predictions. Hu et al. (2018) propose a hybrid attention network to predict the stock trend based on the related sequential news articles. Researchers have also observed the influence of textual data in online media on stock markets (Bollen et al., 2011; Mittermayer and Knolmayer, 2006). Si et al. (2014) showed sentiment analysis based on social media is predictive of each stock's market. However, utilizing multimodal sources of information remains an underexplored avenue in financial forecasting.

Speech Processing and Finance Newer studies (Qin and Yang, 2019; Yang et al., 2020) illustrate the gains obtained using vocal cues from the CEO's earnings conference calls for volatility prediction. Yet, the majority of the current work does not utilize speech based data. The audio features add

greater context and provide psycho-linguistic signaling about the speaker's emotional state (Jiang and Pell, 2017). Qin and Yang (2019) illustrated that late fusion of audio and text features from earnings calls could be used to forecast stock volatility following the earnings call. The verbose quarterly earnings calls (Wang and Hua, 2014) act as a medium of voluntary disclosure (Tasker, 1998), thereby resulting in significant stock movements (Ding et al., 2015), yet the majority of existing approaches do not focus on such highly volatile macro activities, where the market microstructure is highly uncertain (Rogers et al., 2009). During these macro events, the stock returns' predictability can be improved since the disclosure of informed investors influences volatility spreads (Atilgan, 2014). Although multiple sources of information are crucial, not all modalities contribute equally (Akhtar et al., 2019). Noise in one modality can be detrimental in such multimodal frameworks (Morris-Drake et al., 2016).

Multimodality and Finance The Efficient Market Hypothesis (Malkiel, 2003) illustrates the success of multimodal data sources for predictive financial tasks. The more recent multimodal HTML (Yang et al., 2020) is a transformer-based model that uses BERT (Devlin et al., 2019) for textual modeling, and the same hand-crafted audio features as MDRM, in an early fusion formulation. Both MDRM and HTML assume stocks' independence and do not exploit these relations between stock movements. Relations like the same industrial base and co-ownership also result in related stock movements (Feng et al., 2019). Recent works exploit stock relations through graph neural networks (Kipf and Welling, 2017; Veličković et al., 2018) for stock movement prediction (Kim et al., 2019; Sawhney et al., 2020).

Building on these gaps in existing literature, we propose VolTAGE for volatility prediction.

3 Forecasting Stock Volatility

Following Kogan et al. (2009) and Qin and Yang (2019) we define stock volatility as a regression task. For a given stock, with a close price of p_i on trading day i, we calculate the average log volatility over n days following the earnings call as:

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

where, the return price r_i is defined as $\frac{p_i}{p_{i-1}} - 1$ and $\bar{r}$ is the mean of the return price over the period from day-0 to day-n. Additionally, for our auxiliary task we define the single-day log volatility using the daily log absolute returns as follows:

$$v_n = \ln\left(\left|\frac{p_n - p_{n-1}}{p_{n-1}}\right|\right) \quad (2)$$

Problem Statement Given an earning call e, comprising of an audio A, and aligned text T, and stock prices $p_{[0,n]}$, we aim to learn a predictive regression function $f(e_{\{T,A\}}) \rightarrow v_{[0,n]}$.

4 VolTAGE: Architecture and Learning

Below, we describe both the individual components and joint optimization of VolTAGE, and present an overview of the architecture in Figure 1.

4.1 Verbal Cues: Transcript Encoding

We use FinBERT[1] (Araci, 2019) as a sentence encoder, which is a pre-trained language model based on BERT, for language modeling specific to the financial domain. Recent works (Araci, 2019; Keith and Stent, 2019) in this domain indicate the benefits of using a language model pre-trained on financial corpora and retrofitting pre-computed embeddings, achieving considerable performance gains; thereby giving us a strong ground to incorporate the same. FinBERT has been pre-trained on 46,000 documents of financial news articles and has shown state-of-the-art performance on FiQA[2] and Financial PhraseBank benchmarks (Malo et al., 2013).

Formally, we represent the transcript utterances of each call as $(t_1, t_2, ..., t_K)$, where t_i is the i^{th} text utterance and K is the number of sentences, which are encoded as follows:

$$s_i = \text{FinBERT}(t_i) \quad (3)$$

We then pass the sequence of these sentence representations to a BiLSTM as:

$$\overrightarrow{T_t^{(f)}} = \text{BiLSTM}^{(f)}(s_t, T_{t-1}^{(f)}) \quad (4)$$

$$\overleftarrow{T_t^{(b)}} = \text{BiLSTM}^{(b)}(s_t, T_{t+1}^{(b)}) \quad (5)$$

$$T_t = [\overrightarrow{T_t^{(f)}}, \overleftarrow{T_{T-t}^{(b)}}] \quad (6)$$

[1] https://github.com/abhijeet3922/finbert_embedding

[2] https://sites.google.com/view/fiqa

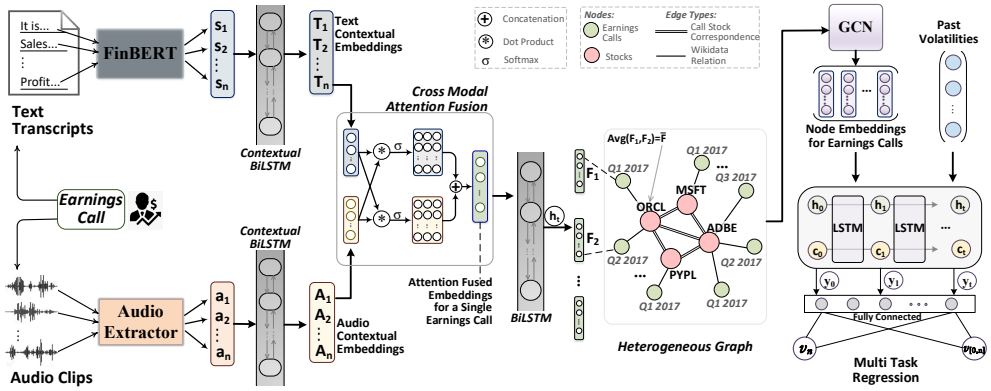


Figure 1: VolTAGE architecture overview: feature extraction, semi-supervised learning and multi-task regression.

4.2 Vocal Cues: Audio Call Encoding

Audio-based features provide prosodic cues related to the affective state of speakers (Montacié and Caraty, 2018). Capturing the emotional valence of the CEO can alter the understanding of the underlying linguistic utterances in an earnings call (Schröder et al., 2001). We extract a set of 26 acoustic features from each aligned audio clip at a sampling rate of $10ms$ for each sentence. These feature time series were then summarized by statistical functions such as mean, median, min, and max to yield a fixed dimensional representation for each sentence. We extend the feature sets of previous works (Qin and Yang, 2019; Yang et al., 2020). These features have shown to be correlated to the speaker's affective states such as stress and anxiety (APQ 11 Shimmer, DDA Shimmer) (Li et al., 2007; Mongia and Sharma, 2014), vocal pace reflecting inconsistencies in vocal cues (ratio of voiced to unvoiced frames in audio) (Přibil and Přibilová, 2009; Viswanathan et al., 2012) and deception (pitch) (Burgoon et al., 2015). We extracted these 26 features from each audio utterance using Praat (Boersma and Van Heuven, 2001).

Text-Audio Alignment Following Qin and Yang (2019), we use the pre-aligned dataset for earnings calls, where the audio is segmented and aligned with each corresponding utterance of the transcript using the Iterative Forced Alignment (IFA) algorithm. IFA is the process of determining the time interval (in the audio file) containing the spoken text for each fragment of the transcript. Qin and Yang (2019) implemented IFA using Aeneas[3] as the fundamental forced alignment method. Formally, we represent the segmented audio clips as $(a_1, a_2, ..., a_K)$ where a_i is the i^{th} audio clip and K being the number of clips of an earning call, with each clip being represented by 26 acoustic features. Similar to the processing of verbal utterances, we employ a BiLSTM layer to sequentially encodes these features, and obtain an audio encoding A_t as:

$$\overrightarrow{A_t^{(f)}} = \text{BiLSTM}^{(f)}(a_t, A_{t-1}^{(f)}) \quad (7)$$

$$\overleftarrow{A_t^{(b)}} = \text{BiLSTM}^{(b)}(a_t, A_{t+1}^{(b)}) \quad (8)$$

$$A_t = [\overrightarrow{A_t^{(f)}}, \overleftarrow{A_{T-t}^{(b)}}] \quad (9)$$

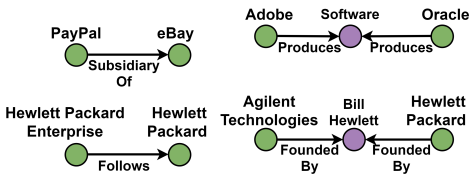
4.3 Verbal-Vocal Attention

The acoustic features provide context and structure to the verbal cues. To capture the associations between verbal and vocal cues, we employ a Cross-Modal Gated Attention Fusion (CM Attn) mechanism that simultaneously learns the alignment weights between audio features and text sentence sequences. Thus, we employ this mechanism to highlight the contributing features by giving more attention to the respective utterance and neighboring utterances. Motivated by Akhtar et al. (2019); Dhingra et al. (2016), we employ the multiplicative gated attention mechanism to generate modality-specific attentive representations.

Formally, a multiplicative gating mechanism is used to attend the important components of text and audio sequences to get the final attentive feature embeddings F_t, F_a which are then combined as:

$$W_a = softmax(T \cdot A^T), W_t = softmax(A \cdot T^T) \quad (10)$$

[3] https://github.com/readbeyond/aeneas



(a) First-order relation (b) Second-order relation

Figure 2: Wiki-company based relations

$$G_a = W_a \cdot T,\ G_t = W_t \cdot A \quad (11)$$

$$H = F_a \oplus F_t = (G_a \odot A) \oplus (G_t \odot T) \quad (12)$$

$\cdot$ is the dot product, $\odot$ represents element-wise multiplication, and $\oplus$ represents concatenation. The fused verbal-vocal feature vector per earnings call is then fed to a GCN, as described next.

4.4 Graph-based Semi Supervised Learning

Mining Stock Relations First, we construct a company graph, inspired by the relations defined by Feng et al. (2019). We mine connections between companies from Wikidata (Vrandečić and Krötzsch, 2014). Wikidata represents relations in the form of statements like (subject; predicate ;object), such as (Facebook; founded by; Mark Zuckerberg).[4] We say that Company A has a first-order relation with company B if there is a statement with A as the subject and B as the object. Similarly, there exists a second-order relation between them if they are related by an intermediate entity. This Wiki-Company graph $G_{WC} = (V, E_{WC})$ is a homogeneous graph, where each node represents a company, and two nodes are connected by an edge representing either a first or second-order relation. We present examples of first and second-order relations in Figure 2. Since the companies are related and not the earnings calls, we extend the graph G_{WC} by incorporating nodes corresponding to earnings calls. Each call is connected to the company it corresponds to through an edge. This extended graph $G(V, E)$ is heterogeneous with two types of nodes (companies and earnings calls).

Graph Convolution Network We frame the task as a graph-based semi-supervised learning problem since we have labels (volatility values) available for a subset of nodes (i.e., earnings call nodes) (Kipf and Welling, 2017). Our intuition behind applying GCNs is to allow the model to distribute gradient information from the supervised loss on the labeled earnings call nodes. As shown in Figure 1, we feed the fused verbal-vocal features H as node features for each earnings call node to the GCN. As for the stock nodes, since a stock may have multiple earnings calls, we consider the mean of feature vectors of all calls pertaining to a stock as its feature vector, to incorporate features across all earnings calls corresponding to that stock. Formally, let $F \in \mathbb{R}^{n \times m}$ represent the input feature matrix comprising these feature vectors of length m for the nodes in G, and D represent the diagonal degree matrix defined as $D_{ii} = \sum_j A_{ij}$. The update rule at layer l of the GCN is then:

$$O^{(l)} = ReLU(\widetilde{A}O^{(l-1)}W^{(l)}) \quad (13)$$

where the first layer is represented as:

$$O^{(1)} = ReLU(\widetilde{A}FW^{(1)}) \quad (14)$$

We experiment with a single layer and a 2-layer GCN, and find better results with the latter. We formulate the exact computation our GCN performs to yield estimated volatility values as follows:

$$O = linear(\widetilde{A}ReLU(\widetilde{A}FW^{(1)})W^{(2)}) \quad (15)$$

Using the earnings call node labels, we train the GCN on the MSE loss using the semi-supervised learning mechanism. This mechanism generates feature representations for both the company nodes and the earnings call nodes, of which we use the latter. Subsequently, these earnings call node features, denoted by O_e are fed along with the financial features to a multimodal LSTM network in a multi-task learning setup as described next.

4.5 Multimodal LSTM for Risk Forecasting

Prior work (Figlewski, 1994) in the financial domain has shown the benefits of using past data for future volatility forecasting. However, fusing the sequential historical volatility data with non-temporal GCN embeddings poses a challenge. To overcome this disparity, we employ multimodal "conditioned" LSTM networks (Karpathy and Fei-Fei, 2015). In our case, we add GCN node embeddings from the first layer with the ReLU non-linearity to the hidden state of the LSTM model at the first time-step to integrate temporally diverse

[4] https://www.mediawiki.org/wiki/Wikibase/DataModel/JSON

modalities. Further, the past data introduces historical context in cases where calls may not have major announcements that would lead to large fluctuations in stock volatility.

Network Optimization To incorporate financial data, we extract past n-day average volatilities prior to the earning call, where $n \in [2, 30]$. Formally, training the LSTM model takes the sequence input vectors $(x_1, .., x_T)$ representing the past financial data along with the earnings call node embeddings O_e, obtained using GCN. The model computes a series of hidden states $(h_1, .., h_T)$ and a sequence of outputs $(y_1, ..y_T)$, by repeating the following recurrence relation from time $t = 1$ to T:

$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + O_e + b_h) \quad (16)$$

$$y_t = softmax(W_{oh}h_t + b_o) \quad (17)$$

Here, $W_{hx}, W_{hh}, W_{oh}, x_i, b_h, b_o$ are learnable parameters and x_t is the average t-day past volatility. Following Karpathy and Fei-Fei (2015), we feed the GCN embeddings to the LSTM only at the first iteration. We use the output y_T from the last LSTM unit for the final multi-output prediction.

Network Optimization We finally train VolTAGE by optimizing a multi-task loss as:

$$\mathcal{L} = \frac{1}{2n}\left(\mu \sum_i (\hat{y_i} - y_i)^2 + (1-\mu)\sum_j (\hat{y_j} - y_j)^2)\right) \quad (18)$$

Here, $\hat{y_i}$, $\hat{y_j}$ are predicted volatilities and y_i , y_j are true volatilities for the main and auxiliary tasks, respectively. μ is a parameter that controls the relative weight of the loss between the two tasks.

5 Experimental Setup

5.1 Data

We used the S&P 500 2017 Earnings Conference Calls dataset (Qin and Yang, 2019).[5] The dataset consists of 559 earnings call audio recordings and their transcripts for 277 public companies in the S&P 500 index. Each call is segmented into a sequence of audio clips aligned with their corresponding text sentences, as spoken by the Chief Executive Officer (CEO) during the call. We temporally divide the data into train, validation, and test sets in a ratio of 70 : 10 : 20 respectively to ensure future data is not used for forecasting. We extract stock prices for each company using Yahoo Finance[6] from 1 January'17 till 31 December'17. The stock data for 11 earnings calls was not available on Yahoo Finance; hence we excluded these calls from our dataset. Following Qin and Yang (2019); Yang et al. (2020), we experiment with $n \in \{3, 7, 15, 30\}$ days to analyze the performance over both short and long term periods.

5.2 Baselines

We compare VolTAGE with the following methods:

- $\mathbf{V_{past}}$: Following Qin and Yang (2019), we use V_{past}, the average log volatility of the past d days to predict the future d days' average log volatility.
- **bc-LSTM:** We also compare against bc-LSTM (Poria et al., 2017) which extracts the uni-modal features using separate contextual Bi-LSTMs and fuses them.
- **MDRM:** Qin and Yang (2019) extract pre-trained GloVe embeddings and hand-crafted acoustic features that are fed to separate BiLSTMs to get their uni-modal contextual embeddings, which are then fused and fed to a two-layer dense network.
- **HTML:** Yang et al. (2020) is the state-of-the-art model using WWM-BERT to encode text tokens. HTML makes use of the same audio features as MDRM. The unimodal features are fused and fed to a sentence-level transformer to get the multimodal representations for each call.

5.3 Training Setup

We tune the hyperparameters on the validation mean square error (MSE) to get: dropout $\delta \in [0, 0.8]$, learning rate $\lambda \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, batch size $b \in \{8, 16, 32, 64\}$ and epochs (< 100).

We use FinBERT with default pre-training parameters, which outputs a 768-dimensional embedding for each sentence. The maximum number of audio clips in any call is 520. Hence, we zero-pad the calls that have less than 520 clips for efficient batching. The number of neurons in the time distributed dense layer following the audio and text BiLSTMs is 100. The heterogeneous graph

[5] We were unable to map price data for 11 data points, which were subsequently dropped

[6] https://finance.yahoo.com/

Model	$\overline{\text{MSE}}$	MSE_3	MSE_7	MSE_{15}	MSE_{30}	R^2_3	R^2_7	R^2_{15}	R^2_{30}
Vpast	1.12	2.99	0.83	0.42	0.23				
LSTM	0.75	1.97	0.46	0.32	0.24	0.34	0.44	0.24	-0.02
HAN (Glove)	0.60	1.43	0.46	0.31	0.20	0.52	0.44	0.27	0.14
MDRM (Audio)	0.60	1.41	0.44	0.32	0.22	0.53	0.47	0.25	0.03
MDRM (Text+Audio)	0.58	1.37	0.42	0.30	0.22	0.54	0.49	0.29	0.06
HTML (Text)	0.46	1.18	0.37	**0.15**	**0.13**	0.61	0.55	**0.64**	**0.42**
HTML (Text+Audio)	0.40	0.85	0.35	0.25	0.16	0.72	0.58	0.40	0.32
VolTAGE	**0.31**	**0.63**	**0.29**	0.17	0.14	**0.79**	**0.65**	0.60	0.39

Table 1: n-day volatility MSE and coefficient of determination R^2 for all models. **Bold** represents the best results.

contains 559 nodes for earning calls connected to the 277 interrelated company nodes. The GCN is trained using Pytorch Geometric (Fey and Lenssen, 2019).[7] We use two GCN layers having 200 and 100 units respectively, inter-spaced by ReLU and followed by a single dense layer. The 200 dimensional feature vectors from the first layer of the GCN after the ReLU activation are fed to a 200-unit conditioned LSTM model for multi-task volatility prediction. We optimize VolTAGE using the Adam (Kingma and Ba, 2014) optimizer.

6 Results and Analysis

6.1 Comparative Analysis

We present the volatility prediction performance of VolTAGE and the baselines in Table 1. We report the MSE averaged across 10 different runs for all models for the main task (n-day average prediction). Our choice of using MSE as a comparative metric is motivated by prior work (Qin and Yang, 2019; Yang et al., 2020). Additionally, we also report the coefficient of determination $R^2 = 1 - \frac{MSE}{MSE_{V_{past}}}$, to illustrate the improvements with V_{past}. We observe gains across the multimodal HTML that leverages both text and audio modalities. We ascribe this improvement to the cross-modal attention fusion mechanism, which uses associations between audio and text modalities over each contextual utterance instead of concatenation used in HTML. Moreover, a key limitation of the baselines is the assumption of independence of inter-stock movements. VolTAGE captures the correlations between price movements of related stocks through the GCN, and hence, volatility, amplifying performance. Similar to prior work (Qin and Yang, 2019), Table 1 illustrates that forecasting volatility in the short-term is a more intricate task than long-term. Based on Post Earnings Announcement Drift (PEAD) (Bernard and Thomas, 1989), a documented financial phenomenon, we note that the price fluctuations around earning calls tend to stabilize over long periods. We observe that VolTAGE outperforms the baselines by a large margin in short-term prediction ($n = 3, 7$); however the margin diminishes over longer durations ($n = 30$).

[7]We extract features for nodes using the last layer of verbal-vocal fusion tuned only for average n-day volatility prediction. The verbal-vocal attention fusion was not trained on multi-task loss, and VolTAGE is not trained end-to-end.

Model	$\overline{\text{MSE}}$	MSE_3	MSE_7	MSE_{15}	MSE_{30}
Glove	0.68	0.99	0.67	0.55	0.49
BERT	0.52	0.85	0.50	0.37	0.35
FinBERT	0.49	0.81	0.50	0.35	0.31
Audio	0.53	0.85	0.52	0.41	0.33
Audio+FinBERT (CM Attn)	0.45	0.77	0.47	0.31	0.24
Audio+FinBERT (CM Attn)+GCN	0.37	0.66	0.39	0.23	0.22
VolTAGE	0.31	0.63	0.29	0.17	0.14

Table 2: Ablation Results over model components

6.2 Ablation Study

We observe an improvement across the text modality (T), when compared to the HTML (Text) model (Yang et al., 2020) in Table 1 and Table 2. This performance gain can be attributed to FinBERT, which is trained to handle language tasks in the financial domain, while the sentence-level transformer employed in the HTML (Text) model is a generalized implementation of BERT (Devlin et al., 2019). We also note that representations learned by FinBERT outperform both GloVe (Pennington et al., 2014) and BERT embeddings, reiterating the effectiveness of domain-specific pre-training. Further, we observe from Table 2, that the Audio+FinBERT (CM Attn) model outperform unimodal components, demonstrating the utility of multimodal verbal-vocal cues for volatility prediction. On adding the GCN, we observe a gain of 17.7%, likely due to the GCN's ability to learn correlations between price movements of related stocks that are captured by the company relations. Finally, on introducing the financial modality via

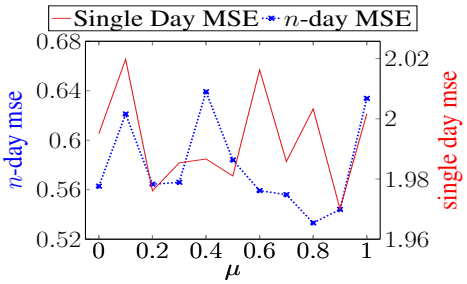
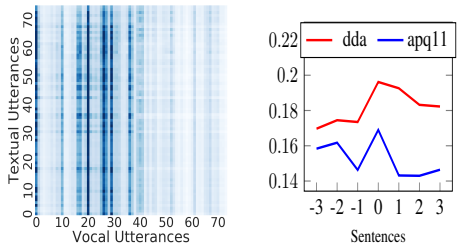


Figure 3: 3-Day Validation MSE vs μ



(a) Audio Aware Text Attention (b) DG: Shimmer Analysis

Figure 4: Verbal and Vocal features from the earnings call for Dollar General from Q3 2017.

the conditioned LSTM network helps in counteracting the impact of PEAD by introducing earnings call independent information into the model; this can be observed in Table 2. We note that VolTAGE outperforms all its ablative components, demonstrating how each of its multimodal components complement each other.

6.3 On Multi-task Learning

Training a network for multiple tasks jointly has shown to improve performance on tasks that share a conceptual similarity (Caruana, 1997). In our case, we optimize VolTAGE on both n-day average and single-day volatility prediction tasks a multi-task formulation. In Figure 3, we analyze the variation of the weight parameter μ with the 3-day validation MSE of n-day average, and single-day predicted volatility. As both tasks share a weighted loss function, by tuning μ, we trade-off between the two tasks. We observe from Figure 3, that at the extreme values of the weight parameter $\mu = 0$ and $\mu = 1$, that represent single task learning on the single-day and n-day average prediction tasks respectively, VolTAGE does not obtain optimal performance. Empirically, we find the optimal $\mu = 0.8$ for 3-day volatility forecasting on the main task, thus validating our hypothesis that multi-task learning across both average and single day spans of volatility prediction improve predictive power.

7 Qualitative Analysis

We analyze the Q3-2017 earnings call for DG (Dollar General), an American variety store company. The stock's price became highly volatile for a few days following the earnings call. Figure 4a shows the audio-aware text attention heatmap for the duration of the earning call. The heatmap represents cross-modal attention weights assigned to textual utterances using corresponding vocal cues. Here each cell (i, j) represents the weight of j^{th} vocal utterance on the i^{th} textual utterance. It is observed that the highest attention is towards the middle of the call, suggesting that the verbal cues of this portion have the highest impact on the text contextual embeddings for most of the sentences in the call. Earning calls of companies are often structured such that the beginning of the call involves introductory disclaimer and greetings, while the CEO starts presenting financial results for the reporting quarter along with future goals of the company towards the middle of the call, which indicates why we see such influential utterances in this portion.

Figure 4b shows the disparity between CEO's vocal and verbal cues around the utterance. While textual content seems positive, a sudden spike in shimmer features in CEO's voice while speaking this sentence suggests disharmony between verbal and vocal cues. Past research in acoustics (Li et al., 2007) suggests an elevated shimmer could be indicative of underlying stress in speech. After the earning call, it was noted that the gross margin of the company slipped by 0.4%, due to the increased transportation costs due to hurricane Irma in 2017. On analyzing the graph, we observe that DG has edge connections with WMT (Wallmart) and TGT (Target Corp.), both of which are retail variety stores, like DG. Analysts had estimated a negative impact of about $2.8 Billion on the retail sector due to the hurricane Irma. This examination is also reflected in the high volatilities recorded for WMT and TGT during the same quarter. A unimodal model may miss these subtle disparities between text and audio. Therefore, VolTAGE, by leveraging cross-modal attention fusion and correlation graphs, accurately forecasts the volatility of DG, three days post the earnings call.

8 Conclusion and Future Work

Volatility, measured as a deviation in returns, is a reliable indicator of market risk linked with a stock. A rich source of company information is earnings calls that provide high risk-reward opportunities given their uniqueness and critical information disclosure. Although evidence shows that enriching models with speech and inter-stock correlations can improve volatility forecasting, this area is underexplored. We propose VolTAGE, a neural architecture that jointly exploits coherence over speech, text, and inter-stock correlations for volatility forecasting following earnings calls. Through experiments on S&P 500 index data, we show the merit of cross-modal gated attention fusion, graph-based learning, and multi-task prediction for volatility forecasting.

There are several promising directions of future work that we wish to explore. First, we want to improve upon the audio feature extraction. To model the speech of CEOs in earnings calls, using semitones rather than raw frequency for pitch-related features. Experimenting with other sets of commonly used acoustic features such as MFCC coefficients, OpenSMILE features and auDeep features for representing audio utterances also form a future direction for audio feature extraction. Second, we want to expand the analysis presented in this paper beyond the S&P 500 index and US-based companies. Existing research (Qin and Yang, 2019; Yang et al., 2020) and this work at the intersection of natural language processing and earnings calls are limited to a small set of companies and earnings calls. Analyzing the demographic, cultural, and gender bias in research pertaining to financial disclosures, particularly earnings calls, forms a future direction of research. We would also want to work on studying a wider set of earnings calls and companies spanning multiple languages, demographics, speakers and gender.

References

Md Shad Akhtar, Dushyant Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task learning for multi-modal emotion recognition and sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 370–379.

Leif BG Andersen. 2007. Efficient simulation of the heston stochastic volatility model. *Available at SSRN 946405*.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Yigit Atilgan. 2014. Volatility spreads and earnings announcement returns. *Journal of Banking & Finance*, 38:205–215.

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 Int*, 5:341–347.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8.

Judee Burgoon, W. Mayew, Justin Giboney, Aaron Elkins, Kevin Moffitt, Bradley Dorn, Michael Byrd, and Lee Spitzley. 2015. Which spoken language markers identify deception in high-stakes settings? evidence from earnings conference calls. *Journal of Language and Social Psychology*, 35.

John Y Campbell, John J Champbell, John W Campbell, Andrew W Lo, Andrew W Lo, and A Craig MacKinlay. 1997. *The econometrics of financial markets*. princeton University press.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Belinda Crawford Camiciottoli. 2011. Ethics and ethos in financial reporting: Analyzing persuasive language in earnings calls. *Business Communication Quarterly*, 74(3):298–312.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. 2016. Gated-attention readers for text comprehension. *arXiv preprint arXiv:1606.01549*.

Ilia D Dichev and Vicki Wei Tang. 2009. Earnings volatility and earnings predictability. *Journal of accounting and Economics*, 47(1-2):160–181.

Francis X Diebold and Kamil Yılmaz. 2014. On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics*, 182(1):119–134.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence*.

Jin-Chuan Duan. 1995. The garch option pricing model. *Mathematical finance*, 5(1):13–32.

Fuli Feng, Xiangnan He, Xiang Wang, Cheng Luo, Yiqun Liu, and Tat-Seng Chua. 2019. Temporal relational ranking for stock prediction. *ACM Transactions on Information Systems*, 37(2):1–30.

Matthias Fey and Jan E. Lenssen. 2019. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Stephen Figlewski. 1994. Forecasting volatility using historical data.

Xue Guo, Hu Zhang, and Tianhai Tian. 2018. Development of stock correlation networks using mutual information and financial big data. *PloS one*, 13:e0195941.

Ehsan Hoseinzade, Saman Haratizadeh, and Arash Khoeini. 2019. U-cnnpred: A universal cnn-based predictor for stock markets.

Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the eleventh ACM international conference on web search and data mining*.

Weiwei Jiang. 2020. Applications of deep learning in stock market prediction: recent progress. *arXiv preprint arXiv:2003.01859*.

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

C Kenneth Jones. 2017. Modern portfolio theory, digital portfolio theory and intertemporal portfolio choice. *American Journal of Industrial and Business Management*, 7:833–854.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Katherine Keith and Amanda Stent. 2019. Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 493–503, Florence, Italy. Association for Computational Linguistics.

Raehyun Kim, Chan Ho So, Minbyul Jeong, Sanghoon Lee, Jinkyu Kim, and Jaewoo Kang. 2019. Hats: A hierarchical graph attention network for stock movement prediction. *arXiv preprint arXiv:1908.07999*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280, Boulder, Colorado. Association for Computational Linguistics.

Werner Kristjanpoller, Anton Fadic, and Marcel C Minutolo. 2014. Volatility forecast using hybrid neural network models. *Expert Systems with Applications*, 41(5):2437–2442.

X. Li, J. Tao, M. T. Johnson, J. Soltis, A. Savage, K. M. Leong, and J. D. Newman. 2007. 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.

Jiexi Liu and Songcan Chen. 2019. *Non-stationary Multivariate Time Series Prediction with Selective Recurrent Neural Networks*, pages 636–649.

Burton G Malkiel. 2003. The efficient market hypothesis and its critics. *Journal of economic perspectives*, 17(1):59–82.

Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. 2013. Financialphrasebank-v1.0.

Marc-Andre Mittermayer and Gerhard F Knolmayer. 2006. Newscats: A news categorization and trading system. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 1002–1007. Ieee.

Saloni Mohan, Sahitya Mullapudi, Sudheer Sammeta, Parag Vijayvergia, and David C Anastasiu. 2019. Stock price prediction using news sentiment analysis. In *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 205–208. IEEE.

Puneet Mongia and R.K. Sharma. 2014. Estimation and statistical analysis of human voice parameters to investigate the influence of psychological stress and to determine the vocal tract transfer function of an individual. *Journal of Computer Networks and Communications*, 2014.

Claude Montacié and Marie-José Caraty. 2018. Vocalic, lexical and prosodic cues for the interspeech 2018 self-assessed affect challenge. In *Interspeech*, pages 541–545.

Amy Morris-Drake, Julie M Kern, and Andrew N Radford. 2016. Cross-modal impacts of anthropogenic noise on information use. *Current Biology*, 26(20):R911–R912.

Mahla Nikou, Gholamreza Mansourfar, and Jamshid Bagherzadeh. 2019. Stock price prediction using deep learning algorithm and its comparison with machine learning algorithms. *Intelligent Systems in Accounting, Finance and Management*, 26.

Yaohao Peng, Pedro Henrique Melo Albuquerque, Jader Martins Camboim de Sá, Ana Julia Akaishi Padula, and Mariana Rosa Montenegro. 2018. The best of two worlds: Forecasting high frequency volatility for cryptocurrencies and traditional currencies with support vector regression. *Expert Systems with Applications*, 97:177–192.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Ser-Huang Poon and Clive WJ Granger. 2003. Forecasting volatility in financial markets: A review. *Journal of economic literature*, 41(2):478–539.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Jiří Přibil and Anna Přibilová. 2009. Spectral flatness analysis for emotional speech synthesis and transformation. *Lecture Notes in Computer Science*, 5641:106–115.

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

Jonathan L Rogers, Douglas J Skinner, and Andrew Van Buskirk. 2009. Earnings guidance and market uncertainty. *Journal of Accounting and Economics*, 48(1):90–109.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. 2020. Spatiotemporal hypergraph convolution network for stock forecasting. In *2020 IEEE International Conference on Data Mining (ICDM)*.

Marc Schröder, Roddy Cowie, Ellen Douglas-Cowie, Machiel Westerdijk, and Stan Gielen. 2001. Acoustic correlates of emotion dimensions in view of speech synthesis. volume 1, pages 87–90.

Jianfeng Si, Arjun Mukherjee, Bing Liu, Sinno Jialin Pan, Qing Li, and Huayi Li. 2014. Exploiting social relations and sentiment for stock prediction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1139–1145.

Jinghua Tan, Jun Wang, Denisa Rinprasertmeechai, Rong Xing, and Qing Li. 2019. A tensor-based elstm model to predict stock price using financial news. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Sarah C Tasker. 1998. Bridging the information gap: Quarterly conference calls as a medium for voluntary disclosure. *Review of Accounting Studies*, 3(1-2):137–167.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

Murlikrishna Viswanathan, Zhen-Xing Zhang, Xue-Wei Tian, and Joon S. Lim. 2012. Emotional-speech recognition using the neuro-fuzzy network. In *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication*, ICUIMC '12, New York, NY, USA. Association for Computing Machinery.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57:78–85.

William Yang Wang and Zhenhao Hua. 2014. A semiparametric Gaussian copula regression model for predicting financial risks from earnings calls. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1155–1165, Baltimore, Maryland. Association for Computational Linguistics.

Linyi Yang, Tin Lok James Ng, Barry Smyth, and Riuhai Dong. 2020. Html: Hierarchical transformer-based multi-task learning for volatility prediction. In *Proceedings of The Web Conference 2020*, pages 441–451.

Jie Zheng, Andi Xia, Lin Shao, Tao Wan, and Zengchang Qin. 2019. Stock volatility prediction based on self-attention networks with social information. In *2019 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, pages 1–7. IEEE.

A Appendices

A.1 Pitch Analysis

We extract the following audio features corresponding to pitch:

1. **Minimum pitch:** The minimum pitch frequency of the frames within a specified time duration.

2. **Maximum pitch:** The maximum pitch frequency of the frames within a specified time duration.

3. **meanF0:** The mean of the fundamental frequency (f0) within a specified time duration.

4. **stdevF0:** The standard deviation of the fundamental frequency (f0) within a specified time duration.

5. **Number of pulses:** The number of pulses within a given time window.

6. **Number of periods**: The number of periods/cycles within a given time window.

7. **Degree of voice breaks:** This is the total duration of the breaks between the voiced parts of the signal, divided by the total duration of the analysed part of the signal.

8. **voiced_frames:** The number of voiced frames. A frame is regarded as locally unvoiced if it has a voicing strength below the voicing threshold (whose standard value is 0.45), or a local peak below the silence threshold (whose standard value is 0.03).

9. **Voiced to total ratio:** The number of voiced frames in a window divided by the total number of frames.

10. **Voiced_to_unvoiced_ratio:** The number of voiced frames in a window divided by the number of unvoiced frames (unvoiced frames are given by total frames-voiced frames).

A.2 Voice Analysis

Under voice analysis, we extract different features quantifying *jitter* and *shimmer* in the earnings call audio. *Jitter* is the relative average vocal perturbation while *Shimmer* is the moment-to-moment amplitude variation. We now describe the various features extracted in this category:

1. **Jitter (local):**This is the fraction of average absolute difference between consecutive periods by the average period.

2. **Jitter (local, absolute):** This is the average absolute difference between consecutive periods in seconds.

3. **Jitter (RAP):** This is the Relative Average Perturbation, the average absolute difference between a period and the average of it and its two neighbours, divided by the average period.

4. **Jitter (ppq5):** This is the five-point Period Perturbation Quotient, the average absolute difference between a period and the average of it and its four closest neighbours, divided by the average period.

5. **Jitter (ddp):** This is the average absolute difference in jitter between consecutive periods divided by the average period. This is Praat's original Get jitter and is proportional to three times RAP.

6. **Shimmer (local) :** This is the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude.

7. **Shimmer (local, dB):** This is the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20.

8. **Shimmer (apq3):** This is the three-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbours, divided by the average amplitude.

9. **Shimmer (apq5):** This is the five-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its four closest neighbours, divided by the average amplitude.

10. **Shimmer (apq11):** This is the 11-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its ten closest neighbours, divided by the average amplitude.

11. **Shimmer (ddp):** This is the average absolute difference of shimmer between the amplitudes of consecutive periods. This is Praat's original

Get shimmer and its value is proportional to three times APQ3.

A.3 Intensity Analysis

We extract the following intensity features:

1. **Mean Intensity:** The mean (in dB) of the intensity values of the frames within a specified time duration.
2. **Minimum intensity:** The minimum (in dB) of the intensity values of the frames within a specified time duration.
3. **Maximum intensity:** The maximum (in dB) of the intensity values of the frames within a specified time duration.
4. **SD_energy:** Standard deviation of energy in the frames within a specified time duration.

A.4 Harmonicity Analysis

We extract the *Harmonics-to-Noise Ratio (HNR)* of the earnings calls audio which has shown to be a measure of the "hoarseness of a speaker".

1. **Harmonics-to-Noise Ratio (HNR):** It represents the degree of acoustic periodicity. It is expressed in decibels. It can be used as a measure for the signal-to-noise ratio of periodic voice signals.