university of groningen

campus fryslân

Exploratory Analysis of Correlation between Earnings Call Acoustic Features and Credit Ratings

A FinBERT Validation Approach

Tiantian Zhang

June 11, 2025

university of groningen

campus fryslân

University of Groningen - Campus Fryslân

Exploratory Analysis of Correlation between Earnings Call Acoustic Features and Credit Ratings: A FinBERT Validation Approach

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Asst. Prof. Dr. Joshua Schäuble (Voice Technology, University of Groningen)
with the second reader being
Supervisor 2's title and name (Voice Technology, University of Groningen)

Tiantian Zhang (S-6055702)

June 11, 2025

Acknowledgements

I hereby express my gratitude to the MSc Voice Technology faculty for their academic guidance. Without their generous sharing of the knowledge and dedication towards students, I would not have been equipped to conduct this study. I am grateful for Ms. Ana Alves, Mr. Corwin de Wolff, and Ms. Abby Toth's guidance on GDPR issues. I also thank the maintenance team for the Hábrók HPC Cluster, which has been instrumental in this education experience.

It was my pleasure to study with a group of brilliant peers. Jan and Stella constantly inspired me with their intelligence and perseverance. I would like to give special thanks to Conny, whose kindness and lively spirit added much joy to my days in the Netherlands.

To baba and mama.
献给爸爸妈妈。

Abstract

This thesis introduces the first systematic investigation of correlation between earnings call acoustic features and companies' subsequent credit rating outcomes by S&P Global, Moody's, and Fitch. It develops an innovative framework that bridges speech technology and corporate financial communication analysis using NLP and machine learning approaches.

While prior research has established relationships between speech sentiments and securities analysts' ratings, no studies have examined correlations with international credit ratings - a critical gap given credit ratings' role in global debt capital markets. This study addresses this intersection using the public Earnings-21 dataset, analyzing earnings calls of 44 US-listed companies, including 24 who received subsequent rating actions (21 affirmations, 2 downgrades, 1 upgrade).

The dataset selection required extensive navigation of international legal frameworks, including GDPR compliance for cross-border speech data. Initial attempts to collect proprietary data were systematically explored under fair use doctrine. As explicit consent from data sources was required, the study chose to use publicly available datasets. This regulatory analysis process demonstrates proficiency in compliance requirements essential for research at the intersection of technology and regulated industries.

Recognizing the data scarcity in this emerging field, the study employs percentile ranking, bootstrap confidence intervals, and MAD-based effect estimation, approaches particularly suitable for small and imbalanced sample. The study further uses the finance-domain NLP model FinBERT as a text sentiment validator. This innovative multimodal validation framework addresses the fundamental ambiguity that physiological arousal can stem from either financial optimism or distress.

Acoustic features, particularly fundamental frequency, pause frequency, and jitter, are extracted and normalized using duration-weighted aggregation to address multi-speaker heterogeneity. The validation framework successfully identifies convergent patterns (aligned acoustic-semantic stress) and divergent patterns (acoustic arousal with positive/neutral sentiment), providing interpretable insights despite limited sample size. The single upgrade case exhibits high acoustic variability coupled with notably negative semantic tone, suggesting complex relationships between speech sentiments and financial outcomes.

The study provides an empirical method for integrating multimodal acoustic semantic analysis with financial outcome indicators, while openly acknowledging the limitations imposed by small sample and data imbalance. Future research is recommended to use larger datasets and multimodal fusion mechanisms. The findings highlight this as foundational work toward operational voice analytics for corporate disclosure analysis in credit assessments.

Key innovations: novel application domain (earnings call speech-credit rating correlation), multimodal validation framework without fusion, small-sample robust statistical methodology.

Key words:
earnings calls, speech sentiments, credit ratings, FinBERT, multimodality, speech technology

Contents

1 Introduction 8

1.1 Background and Motivation 8

1.2 Problem Statement and Study Approach 8

1.2.1 Research gap: lack of research examining speech sentiment against international credit ratings 8

1.2.2 Study approach: descriptive exploration with multimodal validation 8

1.3 Research Questions and Hypotheses 9

1.3.1 Primary research question 9

1.3.2 Secondary research question 9

1.3.3 Hypotheses 9

1.4 Thesis Structure Overview 10

2 Literature Review 11

2.1 Search Strategy and Selection Criteria 11

2.2 Acoustic Features and Benchmark in Earnings Calls 12

2.3 Earnings Call Datasets: Availability and Structure 13

2.4 Progress of Finance-domain NLP Models 13

2.5 Validator Module in Multimodal Analysis of Financial Speech 14

2.6 Credit Rating as Financial Outcome Indicator 14

3 Methodology 16

3.1 Rationale for Correlation Exploration in the Face of Statistical Constrains 16

3.1.1 Post-positivist framework for small-sample research 17

3.2 Dataset Description and Preparation 17

3.2.1 Earnings-21 dataset characteristics 17

3.2.2 Credit ratings: consensus-based classification 18

3.3 Acoustic Feature Extraction Framework 20

3.3.1 Acoustic feature selection 20

3.3.2 Call-level aggregation methodology 21

3.3.3 Duration-weighted acoustic profiling 22

3.4 FinBERT Sentiment Analysis Pipeline 22

3.4.1 Necessity of FinBERT sentiment as directional indicator 22

3.4.2 Domain-specific language model rationale 23

3.4.3 Sentiment classification and validation 23

3.4.4 Acoustic-semantic convergence threshold in multimodal validation 23

3.5 Statistical Analysis Approach 23

3.5.1 Percentile ranking methodology 23

3.5.2 Bootstrap confidence intervals 24

3.5.3 Effect size estimation without inference 24

3.6 Ethical Considerations and Data Privacy 24

3.6.1 Data ethics and privacy 24

3.6.2 FAIR principle and implementation 25

3.6.3 Open science practices 25

3.6.4 Bias and fairness . . . 25
3.6.5 Environmental impact . . . 25
3.6.6 Reproducibility and replicability . . . 25

4 Technical Implementation and System Design 26
4.1 System Architecture Overview . . . 26
4.1.1 End-to-end processing pipeline . . . 26
4.1.2 Component integration design . . . 26
4.1.3 Voice technology framework implementation . . . 26
4.2 Acoustic Processing Implementation . . . 26
4.2.1 Audio preprocessing . . . 26
4.2.2 Feature extraction configuration . . . 27
4.2.3 Feature extraction automation . . . 27
4.2.4 Quality control and validation procedures . . . 27
4.3 FinBERT Integration and Sentiment Processing . . . 27
4.3.1 Transcript preprocessing and segmentation . . . 27
4.3.2 Sentiment score generation and calibration . . . 28
4.4 Interactive Dashboard Development . . . 28
4.5 Statistical Analysis Automation . . . 28
4.5.1 Bootstrap methodology implementation . . . 28
4.5.2 Percentile ranking algorithms . . . 28
4.5.3 Robustness and reproducibility . . . 28

5 Results and Analysis 29
5.1 Baseline Distribution Characterization . . . 29
5.1.1 Affirmation cases acoustic baseline establishment . . . 29
5.1.2 Upgrade/downgrade cases variation analysis . . . 31
5.1.3 Pairwise acoustic feature correlations . . . 32
5.1.4 Sector-specific variation analysis . . . 32
5.2 Non-Affirmation Case Studies . . . 33
5.2.1 Individual downgrade case acoustic profiling . . . 33
5.2.2 Upgrade case distinctive patterns . . . 35
5.3 Acoustic-Semantic Correlation Analysis . . . 36
5.3.1 FinBERT sentiment validation results . . . 36
5.3.2 Convergent and divergent pattern identification . . . 37
5.3.3 Acoustic-semantic-credit rating correlations . . . 38
5.3.4 Semantic-semantic correlations . . . 38

6 Discussion 39
6.1 Interpretation of Acoustic Pattern Findings . . . 39
6.1.1 Hypothesis validation against descriptive results . . . 39
6.1.2 Stress indicator coherence across modalities . . . 39
6.1.3 Individual case study implications . . . 40
6.2 Methodological Contributions and Validation . . . 40
6.2.1 Descriptive exploration framework effectiveness . . . 40

6.2.2 Multi-agency rating consensus approach validation 40
6.3 Voice Technology Applications and Industry Implications 41
6.4 Limitations and Critical Reflections 41
6.4.1 Sample size and class imbalance 41
6.4.2 Data coverage limitation 41
6.4.3 Temporal alignment between calls and ratings 41
6.4.4 Speaker and sector heterogeneity 41
6.4.5 Limited acoustic features and reduced multimodality 41
6.4.6 Statistical and methodological constraints 42
6.4.7 Potential annotation and preprocessing issues 42
6.4.8 External validity and generalizability 42
6.4.9 Technological constrains 42
6.4.10 Ethical and privacy considerations 42

7 Conclusion **43**
7.1 Empirical Method and Findings 43
7.2 Future Research Directions 43

References **44**

Appendices **51**
A Earnings Call Datasets Summary 51
B Semantic - semantic correlation 52
C AI Usage Declaration 53

1 Introduction

1.1 Background and Motivation

Speech sentiment analysis with earnings call audio has been used to support financial qualitative analysis relating to stock price movements and financial fraud detection (Sawhney et al. (2020);Q. Lu, Du, Yang, Xu, and Zhao (2025)). However, limited research has addressed the correlation between speech sentiment and mid-long-term financial indicators such as credit ratings.

The scientific motivation for this research is to correlate earnings call acoustic features with international credit ratings issued by S&P Global, Moody's, and Fitch. Recent research revealed positive correlations between earnings call speech emotion labels (positive or negative during statement, questioning or answering sections) and securities analysts-issued ratings (Chen, Han, and Zhou (2023)). Deep learning architecture combining speech emotion recognition with FinBERT-based sentiment analysis suggested managerial emotion is highly predictive for financial distress, benchmarking Altman's Z-score. International credit ratings have not been used as a financial outcome indicator, despite their position as an international debt capital market benchmark.

The social motivation is to involve speech sentiment as an additional qualitative signal for investors and analysts, complementing existing financial metrics (Rai, Rai, Pakkala, and Thejaswi (2024)). So far, speech sentiment - securities ratings correlation has benefit investors and buy-side analysts, such as investment managers. This study correlates speech sentiment with credit ratings to assist sell-side arrangers and debt issuers in bond pricing. On a broader scale, auditors may consider executives' speech as a risk and fraudulent indicator (Hobson, Mayew, Peecher, and Venkatachalam (2017)). Regulators may monitor corporate behavior and reinforce accountable corporate communications and financial disclosure (Sauter and Jungblut (2023)).

1.2 Problem Statement and Study Approach

1.2.1 Research gap: lack of research examining speech sentiment against international credit ratings

Few research have studied the correlation between vocal sentiment in earnings calls and international credit ratings issued by S&P Global, Moody's, and Fitch, despite having explored the relationship with other mid-long-term financial indicators such as securities analysts' ratings. Chen et al. (2023) linked speech sentiment labels to analyst-issued ratings from WIND database and Chinese rating scales. Hajek and Munk (2023) used spectral features to predict financial distress using Altman's Z-score as the ground truth, neither examined correlation with standardized credit ratings, nor used FinBERT sentiment as a semantic validator rather than a direct feature fuser to interpret acoustic signals.

1.2.2 Study approach: descriptive exploration with multimodal validation

The study adopts a descriptive exploration approach due to severe data constraints: the open source dataset of feasible scale (see Section 2.3 and Appendix A) with 24 rated companies and highly imbalanced rating actions (21 affirmations, 2 downgrades, 1 upgrade). Direct predictive multimodal fusion or survival analysis is precluded by extremely small event rates in minority classes. Instead, call-level acoustic features are extracted and normalized, descriptive statistics and percentiles are computed, and FinBERT-based sentiment score is employed to interpret the direction of acoustic stress markers. This approach prioritizes replicability and seek to establish a baseline for future research when larger datasets become available.

1.3 Research Questions and Hypotheses

1.3.1 Primary research question

How do earnings call acoustic features, specifically fundamental frequency (represented by F0 coefficient of variation and F0 standard deviation), pause frequency, and jitter, correlate with subsequent credit rating actions (affirmation/upgrade/downgrade) issued by S&P Global, Moody's, and Fitch?

1.3.2 Secondary research question

How do FinBERT-derived sentiment scores help validate the acoustic feature's indication of stress or optimism reflected by the subsequent credit rating actions?

1.3.3 Hypotheses

Due to the small data size and thus the exploratory nature of the study, the following descriptive hypotheses are proposed:

H1: Earnings call speech preceding credit rating downgrades will exhibit higher pitch variability (measured by F0 coefficient of variation and standard deviation), increased pause frequency, and increased jitter (voice instability) compared to the affirmation baseline, consistent with psychophysiological stress responses reported in the literature.

Expected observation: Downgrade cases will rank above the 80th percentile of the affirmation distribution on F0 variability, pause frequency, and jitter metrics.

H2: Calls preceding upgrades will show higher overall acoustic feature variation and positive FinBERT-based sentiment classification, reflecting optimism or confidence.

Expected observation: Upgrade cases will display both high acoustic variability and high positive sentiment scores, indicating a positive convergent pattern.

H3: There will be observable convergent or divergent patterns between acoustic features and FinBERT sentiment scores:
High acoustic arousal couple with high negative sentiment before downgrades; high acoustic arousal accompanies high positive sentiment before upgrades.

Expected observation: Case study analysis will identify these patterns descriptively, using percentile ranks and qualitative alignment plots.

All hypotheses are evaluated using percentile ranking, bootstrapped confidence intervals, and effect size estimation relative to the affirmation baseline. Group-level stratification is not attempted due to small sample size. Findings are direct phenomenal observations.

1.4 Thesis Structure Overview

The remainder of thesis consists of six chapters: literature review surveys the existing research (Section 2), methodology details descriptive exploration with FinBERT validation (Section 3), technical implementation describes system architecture (Section 4), results present baseline characterization and case studies (Section 5), discussion interprets findings (Section 6), and conclusion summarizes contributions and future directions.

2 Literature Review

2.1 Search Strategy and Selection Criteria

The literature review employs a 2-step search strategy to comprehensively identify relevant literature at the intersection of speech sentiment and financial indicators, especially credit ratings.

Step 1: Use Google Scholar to overcome discipline limitations of venues to collect maximum publications covering both speech technology and finance.

Search string:

> (“speech” OR “speech sentiment” OR “speech emotion” OR “speech technology” OR “vocal feature” OR “acoustic feature” OR “prosody” OR “SER” OR “audio analysis” OR “voice stress” OR “paralinguistics”)
> AND
> (“finance” OR “financial” OR “invest*” OR “credit rating” OR “credit risk” OR “stock” OR “market” OR “economic” OR “performance” OR “earnings call” OR “conference call” OR “corporate disclosure” OR “capital structure” OR “fraud detection” OR “distress prediction” OR “banking” OR “risk forecasting”)
> AND
> (“IEEE Transactions on Audio, Speech, and Language Processing” OR “Speech Communication” OR “Computer Speech and Language” OR “Interspeech” OR “ACL” OR “ICASSP” OR “ASRU” OR “SLT Workshop” OR “Journal of Finance” OR “Journal of Financial Economics” OR “Review of Financial Studies” OR “Management Science” OR “Financial Management” OR “Decision Support Systems” OR “Journal of Banking and Finance” OR “Journal of Corporate Finance” OR “Review of Finance” OR “Accounting Research” OR “European Accounting Review” OR “SSRN” OR “EMNLP” OR “AAAI”)
> AND
> (site:ieee.org OR site:aclweb.org OR site:signalprocessingsociety.org OR site:nips.cc
> OR site:icml.cc OR site:jmlr.org OR site:cv-foundation.org OR site:afajof.org
> OR site:sciencedirect.com OR site:academic.oup.com OR site:onlinelibrary.wiley.com
> OR site:informs.org OR site:springer.com OR site:ssrn.com OR site:aaai.org
> OR site:acm.org OR site:cambridge.org OR site:taylorandfrancis.com
> OR site:emerald.com OR site:nature.com OR site:annualreviews.org)

Rank the results by relevance and citation volume. No timeline limitation is imposed at this step. OpenAI API is adopted for summarizing selected articles while critical reviewing is performed on the summaries.

Inclusion Criteria: Peer-reviewed articles.
Exclusion Criteria: Non-peer-reviewed studies.

Step 2: Review literature on on acoustic and text financial sentiments and credit rating.

Search string:

> (“speech sentiment” OR “speech emotion” OR “acoustic feature” OR “voice stress” OR “SER” OR “audio analysis” OR “paralinguistics”)
> AND
> (“credit rating” OR “credit risk” OR “earnings call” OR “conference call” OR “financial distress” OR “fraud detection” OR “stock” OR “market” OR “banking” OR “corporate disclosure”)
> AND
> (“IEEE Transactions on Audio, Speech, and Language Processing” OR “Speech Communication” OR “Interspeech” OR “ACL” OR “ICASSP” OR “Journal of Finance” OR “Journal of Financial Economics” OR “Review of Financial Studies” OR “Management Science” OR “Journal of Banking and Finance” OR “SSRN”)
> AND
> (site:ieee.org OR site:aclweb.org OR site:sciencedirect.com OR site:onlinelibrary.wiley.com OR site:ssrn.com OR site:informs.org OR site:springer.com)

Inclusion Criteria: First-rank venues; published in the last three years.
Exclusion Criteria: Non-first tier venues; non-replicable method; no statistical result disclosure; in case of no code disclosure, no mathematical formula disclosure.

2.2 Acoustic Features and Benchmark in Earnings Calls

Hobson, Mayew, and Venkatachalam (2012) initiated that vocal dissonance can detect financial misreporting. They (Mayew and Venkatachalam (2012)) further expanded this into managerial affective states, stating that positive or negative emotions during earnings calls correlated with future earnings. Baik, Kim, Kim, and Yoon (2024) analyzed 28,515 earnings calls with wav2vec 2.0 and identified key acoustic indicators as F0 shifts, pause patterns, jitter and shimmer, harmonics-to-noise ratios, and speech tempo. M. Miao, Wang, Li, Jiang, and Yang (2024) found speech rate, pitch, and emotionally arousal correlate with crowdfunding success.

Fundamental frequency (F0) has been consistently identified as a main stress indicator (Giddens, Barron, Byrd-Craven, Clark, and Winter (2013)). Rising F0 is interpreted as stress, as research (Broś (2023); Kappen et al. (2022); Kappen, Vanhollebeke, Van Der Donckt, Van Hoecke, and Vanderhasselt (2024)) show enduring F0 volatility during stress-inducing tasks, highlighting its prominence in physiological stress response (Bänziger and Scherer (2005)). In earnings calls, F0 standard deviation (F0_stv) is found significantly correlated with positive emotions, but not negative emotions (Gobl and Chasaide (2003); Johnstone and Scherer (2000); Mayew and Venkatachalam (2012)). If to predict post-earnings call stock volatility, omitting pitch or standard deviation of pitch raise mean squared error by 0.7% and 0.65%, respectively (Qin and Yang (2019)). Greater F0 fluctuation in management’s voice during IPO roadshows indicates positive emotions, and in turn higher first-day stock return (Zhang, Li, He, and Liang (2024)).

Pause frequency increases or lengthens as stress heighten (Trouvain and Grice (1999)). Studies showed higher cognitive load increases spoken hesitation and leads to less fluent speech (Shriberg (2001)).

Jitter (vocal frequency variation) has also been found to be a stress-induced anxiety indicator with high validity (Fuller, Horii, and Conner (1992); Schuller et al. (2014)). Research showed that jitter and shimmer improve emotion classification accuracy when added to baseline spectral and energy features (Eyben, Wöllmer, and Schuller (2010); X. Li et al. (2007); Vlasenko, Schuller, Wendemuth, and Rigoll (2007)). In finance, Jitter local is extracted in vocal cue stock prediction research (Qin and Yang (2019)).

Earnings calls involve multiple executives and analysts and contain both prepared speech and Q&A sessions. Both sessions are speech sentiment informative (Hynes, Garvey, and O'Brien (n.d.)). Analyzing sentiments at the call level has proven effective for predicting stock performance (Cao et al. (2024)).

2.3 Earnings Call Datasets: Availability and Structure

Earnings call datasets vary drastically on scales. SPGISpeech (O'Neill et al. (2021)) released by S&P Global in 2021 is the largest opensource dataset so far, with 5,000 hours of recordings spanning 2007-2020, followed by the MAEC (Multimodal Aligned Earnings Conference Call J. Li, Yang, Smyth, and Dong (2020)) dataset covering 3,443 calls spanning 2015 to 2018 totaling 921 hours of recordings. For academic purposes, Earnings-21 (Del Rio et al. (2021)) is often analyzed for its multimodality (audio and text), meta data (company names, sectors, speakers) availability, all recordings in 2020, and size of 39 hours. While the larger datasets are challenging to annotate and validate for this study especially without credit rating metadata, Earnings-21's small scale disabled multimodal fusion. Middle-sized datasets often remain proprietary (see Appendix A).

Institutional research is dominated by commercial datasets such as FactSet, Refinitiv (LSEG), and S&P Capital IQ. LSEG processes approximately 7,000 companies while Capital IQ covers around 8,000 public companies. These commercial sources provide rich metadata together with structured financial metrics, although subscription ranges $12,000 to $25,000 annually.

2.4 Progress of Finance-domain NLP Models

Benchmarks studies have consistently shown domain-specific models such as FinBERT and BloombergGPT outperform generic models (Shah et al. (2022); Wu et al. (2023)), The landmark, Loughran and McDonald (2011) financial dictionary, revealed 75% of misclassified financial words in Harvard's general sentiment lexicon. FinBERT emerged in 2023 and excel in sentiment analysis with 88.2% sentiment classification accuracy on financial texts, albeit not necessarily optimal for other financial NLP tasks, such as entity recognition, relationship extraction, and numerical reasoning (Alissa and Alzoubi (2022); Kirtac and Germano (2024)).

Controversies revolve around dataset quality vs. quantity (Dang and Verma (2025); R. Verma (2024)), lack of disclosure on confidence intervals, and using inappropriate metrics on sequential data (Wasserstein, Schirm, and Lazar (2019)). Among the essential accuracy drivers are annotation quality (Grosman et al. (2020)), class balance (H. Lu, Ehwerhemuepha, and Rakovski (2022); Tomanek and Hahn (2009)), and proper validation protocol (Cejas, Azeem, Abualhaija, and Briand (2023)). Architecture choice shows moderate but consistent benefits (Lipenkova (2022)). For example, RoBERTa-base models outperform BERT (Astuti and Alamsyah (2024)) by 3-7% on financial text (J. Miao, Lin, Luo, and Liu (2024)), but RoBERTa-large only gains 2-4% accuracy compared to the base model (Liao and Shi (2022)).

Domain-specific pretraining provides significant empirical benefits for accuracy gain (15-17% Araci (2019)) . Financial PhraseBank (Malo, Sinha, Korhonen, and Wallenius (2014)) dataset for instance provides 4,840 sentences annotated from investor's perspective rather than general sentiments. This study focuses on SEC-listed companies' earnings calls, and the model choice relies on the pretrain dataset. FinBERT which was pretrained with corporate annual and quarterly filings from SEC's EDGAR website, over 476k financial analysts' reports issued for S&P firms, and 136k earnings call conference scripts (Huang, Wang, and Yang (2023)) emerged as the most suited model for this study.

2.5 Validator Module in Multimodal Analysis of Financial Speech

While there is not one architectural consistently outperform others in financial speech multimodal fusion, both cross-modal credibility assessment where one modal serves as validator, and direct fusion with integrated fusion learning have seen progress. Kaikaus, Hobson, and Brunner (2022) used bidirectional LSTM to validate text sentiment with acoustic features. The validation approach offers high interpretability for regulatory compliance, clear construction validity, and the ability to identify specific inconsistencies between modals (Hennig, Firk, and Wolff (2025)).

On the other hand, the direct fusion combines acoustic and semantic features at the feature level. Mathur, Goyal, et al. (2022)'s DocFin architecture achieved 5-12% accuracy gain in stock price movement prediction by integrating tabular financial data with multimodal earnings call data.

The distinction between the two approaches is compared by Throckmorton, Mayew, Venkatachalam, and Collins (2015) who found fusion more accurate for fraud prediction, but only with robust feature selection approaches. The controversy deepened by skepticism over LVA being used for validation, raising concern over detectability of genuine emotion states (Maniar, Rathod, Kumar, and Jain (2022)).

The field continues to progress towards large language model integration with acoustic analysis, providing more sophisticated semantic information, real-time processing (Baik, Kim, Kim, and Yoon (2023)) capacities for live earnings calls to enable market applications (Doran, Peterson, and Price (2012); Froot, Kang, Ozik, and Sadka (2017)), as well as standardized evaluation protocols (J. Li et al. (2020)).

2.6 Credit Rating as Financial Outcome Indicator

Empirical use of credit ratings as dependent variables traces to Kisgen (2006) who established that companies systematically adjust their capital structure decisions near rating changes.

Sangiorgi and Spatt (2017) further demonstrated that regulatory reliance on credit ratings create mechanical relationship towards firms' cost of capital. However, conducting research on credit ratings involve various methodological considerations. The interpretation of rating categories varies by agencies, leading to potential inconsistency in empirical research (Charlin and Cifuentes (2017); Matthies (2013)). Sample selection requires controlling issuers, industries, and time periods factors as ratings are influenced by sectoral risks (Lopatta, Tchikov, and Körner (2013)). Statistical method is another critical consideration. While traditional methods such as logistic regressions have been widely used to model the relationship between credit ratings and various predictors (Zhao et al. (2015)), recent research has adopted multilayer perceptron and classification and regression trees to improve the accuracy of credit rating prediction (Overes and Van Der Wel (2023)).

Consultation papers have discussed that positive credit-related sentiments in earnings calls predict favorable rating actions and lower credit default swap spreads. There has been limited research discussing acoustic feature correlation with credit ratings. Research typically conducts binary classification of credit ratings (investment vs. speculative grade) (Brown, Chen, and Kim (2015)), ordinal models capturing notch changes (Berteloot et al. (2013)), and continuous proxies such as credit spreads (Hirk, Hornik, and Vana (2019)).

Implementing credit rating research requires navigating complex data access issues. Limited database is accessible for academic use (e.g., WRDS's Compustat, CRSP, etc.). Sorting from the agency website pose selection and survivorship bias, requesting researchers to use survivorship bias-free samples and consider Heckman-type selection (Toomet and Henningsen (2008)) corrections for non-random rating coverage.

The integration of textual analysis with credit ratings involves topic modelling through Latent Dirichlet Allocation (Loughran and Mcdonald (2020)) to identify risk factors in disclosure, word embedding to capture contextual meaning (Hlongwane, Ramaboa, and Mongwe (2024)), and graph neural network constructing corporate similarity networks from SEC filings (Das, Huang, Adeshina, Yang, and Bachega (2023)). Slapnik and Lončarski (2023) identified the qualitative judgement of rating committee in sovereign ratings by extending traditional regression with new measures obtained from textual sentiment analysis.

For speech sentiment analysis, researchers must control for disclosure characteristics such as report length, readability indices, and frequency of forward-looking statements. Using credit ratings (Partnoy (2017)) for financial communication research requires acknowledgement of agency bias and temporal inconsistencies.

3 Methodology

3.1 Rationale for Correlation Exploration in the Face of Statistical Constrains

While addressing the gap of limited precedent in using credit ratings as financial outcome indicator, this study must situate itself in severe data constrains. Current opensource earnings call datasets have polarized scales (see Appendix A). Choosing certain calls from SPGISpeech (5,000 hours) or MAEC (921 hours), even if annotated the absent credit ratings and controlled the segment distribution, confounding issues persist as the calls span 2007 to 2020, suggesting speech sentiments are potentially swayed by macroeconomic environment (e.g., 2008 financial crisis and 2020 covid-19). Earnings-21 was used as its scale (44 calls) enabled fine-grained credit rating annotation from three agencies and validation, and all calls were from 2020, controlling macroeconomic factor.

However, this data limitation prevented the study from conducting direct acoustic feature vs. credit rating correlation analysis, as only 24 companies in the datasets were rated by S&P Global, Moody's, and Fitch with disparate rating action grouping (21 affirmation, 2 downgrade, 1 upgrade). Excluding the credit ratings from the study scope would remove the ground truth indicator with the marginal benefit of 20 more calls, compared to 848 calls in Chen et al. (2023)'s research and 1278 calls in Hajek and Munk (2023)'s. Although time gap between earnings call dates and subsequent rating action dates could potentially enable survival analysis, the only 2 or 1 event in the minority classes disabled this approach. The time gap ranging 14-606 days compounded this issue, rendering cross-validation meaningless.

Inconsistent availability of speaker roles restricted speaker-specific analysis, as annotation from public sources remained arbitrary and sometimes impossible (certain calls did not disclose speaker names). Sectorial stratification faced similar statistical power issues. The total 44 companies are segmented into 9 industry sectors (3 to 6 companies per sector), reducing statistical efficiency and complicating interpretation, despite successful Interspeech precedents operate on small datasets such as low-resource languages (J. Wang, Zhu, Fan, Chu, and Alwan (2021); Zhong et al. (2022)).

Refraining from multimodal fusion with 24 hours of rated companies' audio, the study adopted FinBERT as a directional validator to acoustic stress indications, addressing the fundamental ambiguity that physiological arousal can steam from financial optimism or distress. Without semantic context, elevated acoustic features remain uninterpretable.

The study eventually took on a descriptive exploration approach, extracted the selected acoustic features at the call-level, normalized features within each company, computed descriptive statistics and percentiles, applied FinBERT for sentiment validation, correlated acoustic and semantic features, and performed case study profiling for non-affirmation rating cases. It prioritized transparent disclosure of limitations and replicable approach, while sought to leave an empirical record of acoustic feature - credit rating correlation exploration as a reference for future research supported by larger data. These methodological decisions align with current best practices for small-sample, exploratory research in computational paralinguistics and financial analytics.

3.1.1 Post-positivist framework for small-sample research

As an exploratory analysis of the communication climate within an organization, call-level aggregation is justified (Patterson et al. (2005)). Deterministic claims are precluded given credit rating's structured, multi-factor evaluations that integrate financial metrics, regulatory frameworks, and qualitative judgments. This strategy is supported by recent work in missing data theory (Graham and Graham (2012)), which shows that when ground truth labels are intrinsically noisy, direct acoustic-outcome correlations offer more interpretable baselines than multimodal fusion.

3.2 Dataset Description and Preparation

3.2.1 Earnings-21 dataset characteristics

The Earnings-21 dataset contains 44 public earnings calls from 44 distinct companies recorded throughout 2020. The sampling rates range from 11,025 Hz to 44,100 Hz. The total duration is 39 hours and 15 minutes, and each recording ranges from 17 minutes to 1 hour and 34 minutes. There are 2 to 20 speakers on each recording. The dataset includes a variety of speaker profiles, including professional roles (e.g., C-suite members, financial analysts, operators, etc.), speech styles (e.g., verbally spontaneous with disfluencies, natural conversation style, technical jargon-rich, etc.), and metadata that records the names of the speakers and the companies. The dataset, which covers nine industry sectors - Basic Materials, Conglomerates, Consumer Goods, Financial, Healthcare, Industrial Goods, Services, Technology, and Utilities-offers a thorough depiction of corporate communication in major sectors. To determine the time gap with the subsequent rating action dates, call dates were added with reference to the dataset source, Seeking Alpha.

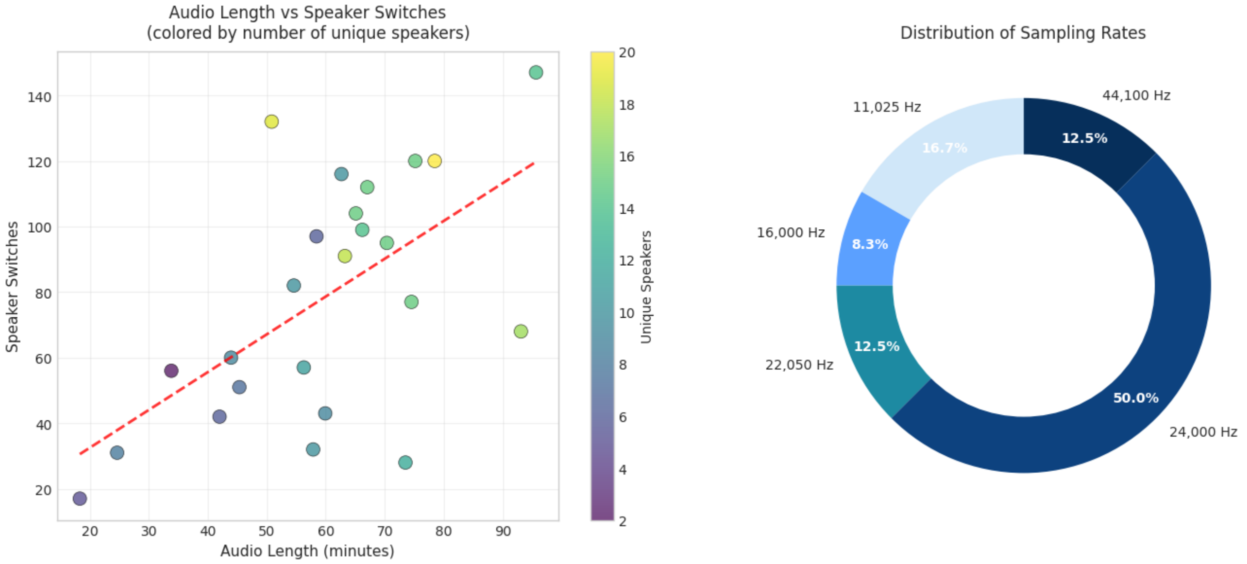


Figure 1: Earnings-21 dataset characteristics

No Covid-19 related macroeconomic controls (e.g., ”stable/deteriorating” sectors in 2020) are implemented since both speech features and credit ratings reflect contemporaneous company performance and inherently reflect the impact of Covid-19. Also, any cross-company comparison within each sector accounted for the shared macroeconomic context, making additional Covid-specific adjustments unnecessary.

Professional roles are not provided in the dataset. Number of speakers per call was considered in the case study. To comply with privacy-preserving research practices and GDPR Article 4(1) requirements for protecting identifiable natural persons even in publicly available contexts, individual speaker names were pseudonymized, while company names remained unchanged as they represent public legal entities outside the GDPR's natural person scope.

3.2.2 Credit ratings: consensus-based classification

Financial scientific literature consistently uses consensus or average ratings across agencies to reduce measurement error or agency-specific biases (Lehmann and Tillich (2014)). Studies demonstrate that composite ratings provide more stable and predictive measures than single-agency ratings. Corporate finance research treats rating disagreements as measurement uncertainty rather than conflicting truths, supporting composite classification approaches (Norden and Roscovan (2014)).

24 companies received first-time or surveillance public ratings by at least one of S&P Global, Moody's and Fitch after the earning calls. The rest 20 companies were never rated, privately rated, or withdrew their ratings after the calls. For companies rated by more than one agency, upgrades and downgrades are prioritized over affirmations, as such rating actions are more likely to reflect significant changes in company merits or risks, thus more helpful in identifying portions of the earnings call that are more closely correlated with rating movements and more valuable for analyzing sentiment changes. The ratings show primarily coverage disagreements (unrated vs. rated) rather than directional disagreements (upgrade vs. downgrade), suggesting coverage limitations rather than fundamental analytical differences.

For consistent actions by more than one agency, the nearest subsequent rating action is selected to align with the timing of the earnings call sentiments most closely, reflecting a time gap between 14 to 606 days. The time gap is consistent with S&P Global's observation that significant differences in the text sentiment of the rating reports associate with a higher likelihood of rating movements within the following 24 months of the rating reports publication. The variability in the time gap reflects various surveillance schedules across rating agencies and sectors, according to their respective rating criteria and disclosure according to Paragraph (a)(1)(ii)(K) of SEC Rule 17g-7.

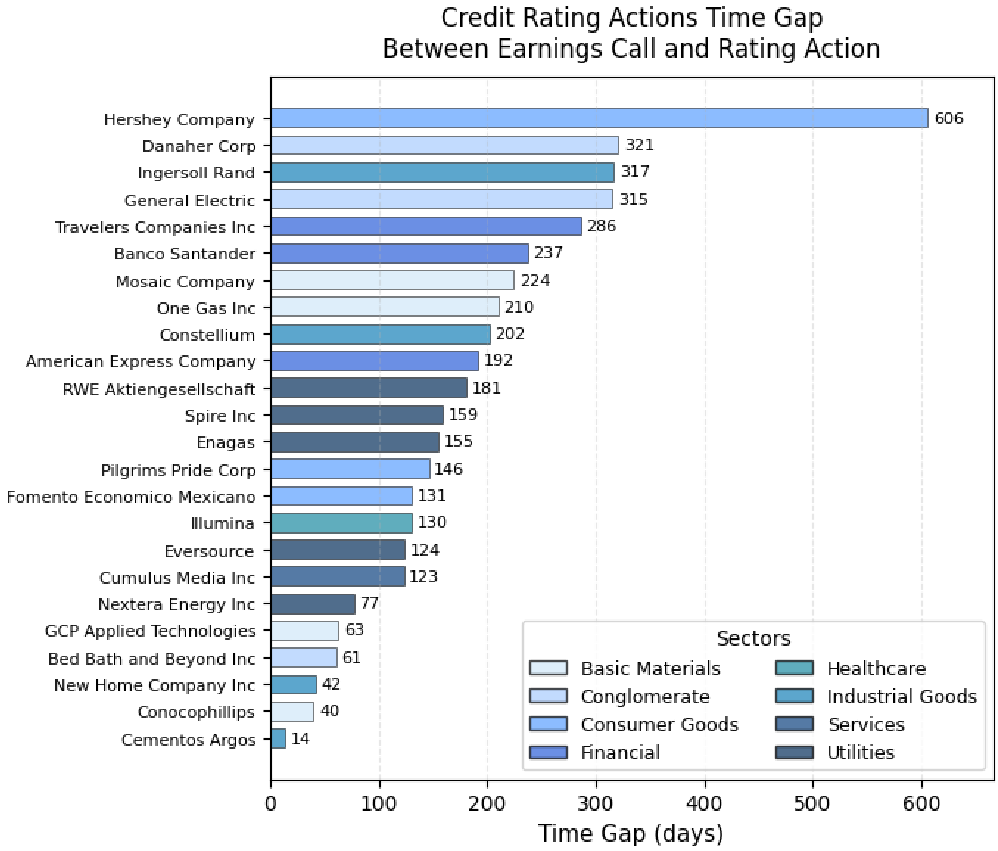


Figure 2: Earnings call and rating action time gap

For companies whose senior unsecured ratings coexisted with equity unit ratings (e.g., Spire Inc. received ‘BBB’ on its equity units from S&P Global, two notches below the ‘A-’ issuer rating 9 months after the earnings call), the senior unsecured rating was considered due to its tighter linkage to the company’s overall credit quality and its priority in the capital structure.

Sectors are considered when comparing the volatility of acoustic features of speeches within each sector. However, further stratification by rating subcategory (e.g., investment grade, speculative grade, etc.) within each sector is not performed due to the small sample size.

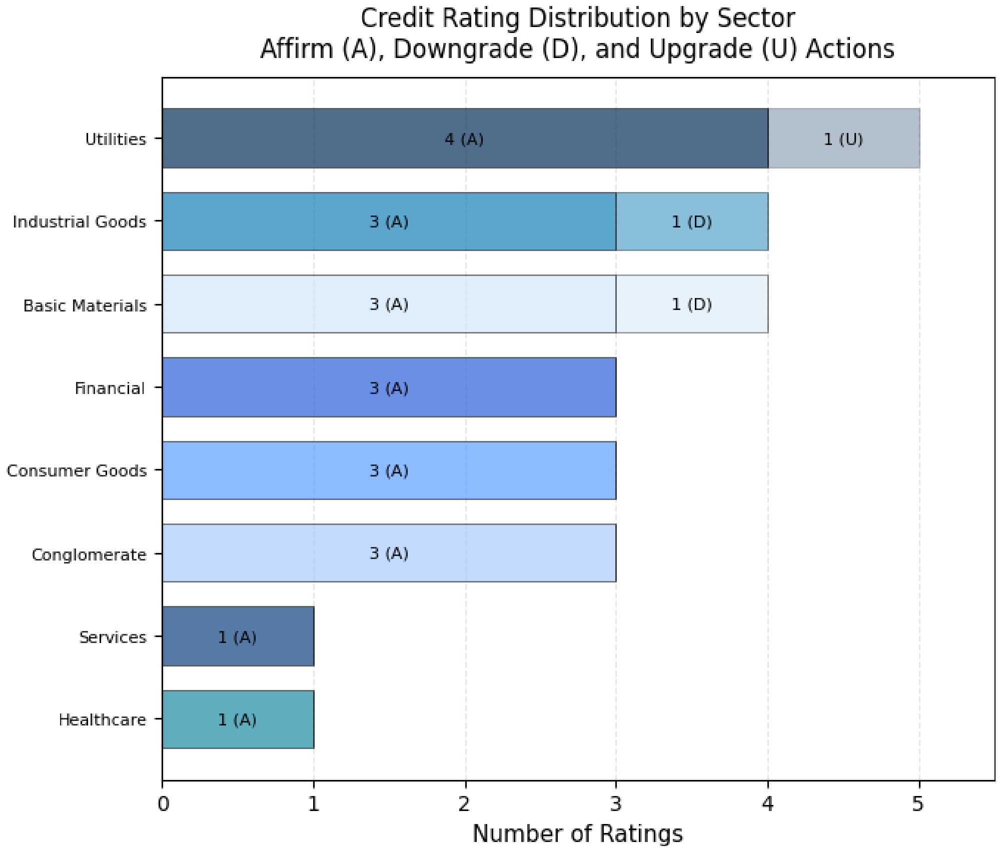


Figure 3: Credit rating distribution by sector

The study uses consensus affirmations (n=21) as baseline distribution for acoustic feature characterization, while analyze consensus downgrades (n=2) and upgrades (n=1) as case studies with percentile ranking against consensus baseline.

Limitations are acknowledged including various rating methodologies across agencies, unrated companies may differ systematically from rated companies in ways that affect earnings call communication patterns, as well as the timing differences between earnings calls and subsequent rating actions across agencies.

3.3 Acoustic Feature Extraction Framework

3.3.1 Acoustic feature selection

The selection of acoustic features for earnings call sentiment analysis is based on literature precedents (M. Miao et al. (2024); Throckmorton et al. (2015)) and psychophysiological stress response theory (Forbes and Pekala (1993)). Fundamental frequency variability (f0_cv) (Haas (2022)) is a key stress indicator given its physiological basis in autonomic nervous system activation. Regardless of semantic content, pitch dynamics can be measured when financial stress activates the sympathetic nervous system (Van Puyvelde, Neyt, Mcglone, and Pattyn (2018)).

The study initially extracted 70 distinct features including 18 for F0, 7 for voice quality (e.g., jitter, shimmer, hnr), 6 temporal features (e.g., speech rate, pause frequency, pause duration, speaking time ratio), 39 spectral features, as well as mean and standard deviation for MFCCs. However, based on Raudys and Jain (1990)'s statistical learning framework for small samples , maximum 8 features are supported by the n=24 dataset for reasonable linear discrimination analysis. As preliminary experiment showed that F0 coefficient of variation (f0_cv), F0 standard deviation (f0_std), pause frequency, and jitter local showed more conclusive results, this study focuses on these four features.

In stress detection tasks across various domains, recent meta-analyses (Mousikou, Strycharczuk, and Rastle (2024)) show that f0 variability performs better than static f0 measures, with effect sizes (Cohen's d = 0.72-0.89) greater than those of other acoustic parameters. Speech rate and pause patterns are complementary temporal features that capture the cognitive load aspects of stress and show how executive function is disrupted during uncertain financial times (O'Neill et al. (2021)). Measurements of jitter offer indicators of voice quality that are sensitive to minute tremors in the patterns of vocal fold vibration under stress (Mahon and Lachman (2022)). By addressing the known drawbacks of unimodal approaches (Skrlj (2024)), this feature selection produces a multidimensional acoustic profile that captures both autonomic arousal and cognitive load dimensions of financial communication stress while preserving measurement reliability under a variety of recording conditions.

3.3.2 Call-level aggregation methodology

Initially, the study explored within-speaker acoustic feature movement in the time domain by segmenting each speaker's speech according to credit rating-relevancy by manually aligning transcript sentences with rating press releases based on credit rating domain knowledge. This approach was abandoned due to irreproducibility and subjectivity. Finance-domain models (e.g., FinBERT) are effective at sentiment or risk tagging, not capable of mapping speech tokens with rating rationale. Iterative, human-in-the-loop annotation in the credit rating context remains underexplored and beyond this study's scope.

The decision to employ call-level acoustic aggregation was informed by the practical constraints of multi-speaker financial discourse as well as recent developments in organizational communication theory (De Benedicto, Sugahara, Silva Filho, and Sousa (2018)). The emergence of collective communication patterns at organizational levels that surpass individual variation results in quantifiable communication climate signatures (P. Verma (2013)).

Recent computational paralinguistics research (Haider, De La Fuente, and Luz (2019)) shows that even with speaker heterogeneity, conversation-level acoustic features can successfully capture group-level phenomena. Thus, this methodology utilized distributional robustness measures. Modern missing data theory (Graham and Graham (2012)) supports this strategy, showing that systematic aggregation yields more insightful measurements than discarding data with incomplete metadata. Several distributional parameters were extracted.

The limitations include loss of speaker-specific information and risks of obscuring intra-call temporal patterns that may be crucial for stress detection. Thus, the methodology present organizational-level patterns as findings rather than definitive conclusions.

3.3.3 Duration-weighted acoustic profiling

A theoretically supported method for reducing speaker heterogeneity in multi-participant financial discourse is duration-weighted acoustic feature extraction (Ji, Hou, Jin, and Li (2013)). This approach is based on frameworks for information asymmetry in financial communication (Ivanitsky and Tatyannikov (2018)). Executive speakers, such as CEOs and CFOs, typically comprise 60-80% of the content of earnings calls and offer disproportionately valuable information about the company's financial health. The importance sampling principles of statistical theory are mathematically implemented by duration weighting (Tokdar and Kass (2010)).

Recent studies that commonly use duration-weighted acoustic features in multi-speaker settings support the approach empirically (Hogg, Evers, Moore, and Naylor (2021)). By minimizing the impact of short, non-executive contributions (analyst questions, operator announcements) that introduce noise rather than signal regarding organizational stress states, duration weighting applies optimal filter to the feature statistics.

3.4 FinBERT Sentiment Analysis Pipeline

3.4.1 Necessity of FinBERT sentiment as directional indicator

Involving FinBERT sentiments as directional indicators in this study was motivated by the statistical constrains faced by a unimodal acoustic feature vs. credit rating correlation approach. The total of only 24 samples with an extreme 21:2:1 class imbalance violates the proportional hazards assumption (Ng'andu (1997)) of survival analysis (Machin, Cheung, and Parmar (2006)), which treats minimum 10-15 events per covariate as a rule of thumb for stable models. The method lacks enough statistical power for trustworthy inference because there are only two or one events in the minority classes while current survival analysis guidelines suggest 200+ events to detect moderate effect sizes. These problems are exacerbated by the 14-606 day temporal gaps, and cross-validation is useless with such small samples. Statistical significance cannot be attained for realistic effect size detection without 8-10 times larger than the available data.

Using FinBERT as a directional validator rather than a fused feature extractor is supported by Cross-Modal Consistency Framework, which was introduced in 2024 research (Liang, Zadeh, and Morency (2024)) and offered mathematical support for validation-based techniques where consistency losses enforce cross-modal alignment without the need for direct fusion. The information bottleneck principle (Tishby and Zaslavsky (2015)) implies that validation techniques can successfully regulate information flow, avoiding the information dilution typical of direct fusion. Auxiliary Task Learning Frameworks (Kumar (2024)) demonstrated how pre-trained models serve as validators through auxiliary losses that enforce cross-modal consistency. By using coordinated representations rather than joint fusion, this method ensures alignment while preserving modality-specific information.

Sentiment analysis models have demonstrated successful validation patterns. In financial NLP, FinBERT has been applied as an auxiliary supervisor rather than a direct predictor (e.g., in quantitative finance models, FinBERT's sentiment scores correct biases in structured-data predictions (Shobayo, Adeyemi-Longe, Popoola, and Ogunleye (2024))). Studies (Du, Xing, Mao, and Cambria (2024)) use FinBERT as a validation mechanism for public confidence indicators in systemic risk prediction.

3.4.2 Domain-specific language model rationale

FinBERT (Huang et al. (2023)) demonstrates superior performance on sentiment classification tasks especially on financial terminology, risk assessment language, and earnings call-specific expressions compared to general models. Trained on SEC filings and financial documentation, FinBERT provides better comprehension of compliance-driven language patterns typical in earnings calls, and thus its sentiment classification shows better alignment with financial analyst assessments than general sentiment models.

3.4.3 Sentiment classification and validation

The study uses FinBERT sentiment classification to provide linguistic support for interpreting acoustic stress markers. Financial distress typically corresponds with negative semantic content, whereas positive events might display divergent patterns. Given that stress and excitement can produce similar acoustic arousal but different semantic valence, the methodology generates a validation matrix wherein: (1) stress appears as acoustic-semantic negativity convergence, (2) optimism exhibits acoustic correlations, and (3) acoustic features and FinBERT-derived sentiment polarity (negative/neutral/positive) are baselined.

3.4.4 Acoustic-semantic convergence threshold in multimodal validation

The study adopts multimodal validation without fusion. The study made the following assumptions, although Cohen (2023)'s benchmarks were r = 0.10 (small), r = 0.30 (medium), and r = 0.50 (large) effect sizes. Pearson's $r \geq 0.6$ indicates strong correlation, as ECB institutional research (Andersson, Neves, and Nunes (2023)) accepts 0.5–0.6 as meaningful earnings call evidence. $0.3 \leq r < 0.6$ is considered moderate, as voice stress research reports correlation between psychological measures and acoustic features ranging from 0.3 to 0.7 (Van Puyvelde et al. (2018)). Finally, $r < 0.3$ is classified as weak correlation.

3.5 Statistical Analysis Approach

3.5.1 Percentile ranking methodology

The percentile ranking method (Bornmann, Leydesdorff, and Mutz (2013)) was selected for its robustness in non-parametric analysis of non-normal acoustic features and idiosyncratic credit rating actions in this small sample, in line with contemporary statistical practices.

Non-parametric percentile ranking evaluates individual cases based on their empirical position relative to a baseline distribution of affirmations (e.g., n=21). For a given feature (e.g., F0 variability), the percentile rank of a downgrade case d is computed using the standard definition:

$$\text{PercentileRank}(d) = \frac{L + 0.5 \times E}{N} \times 100$$

where:

$$\begin{aligned}
d &= \text{value to rank (e.g., for a downgrade case)} \\
L &= \text{number of baseline values less than } d \\
E &= \text{number of baseline values equal to } d \\
N &= \text{total number of baseline cases}
\end{aligned}$$

This rank reflects how extreme d is within the baseline - e.g., a value of 95% indicates that d is greater than 95% of affirmation cases, accounting for ties. This empirical approach makes no distributional assumptions (e.g., normality), making it suited for small-sample settings, where parametric assumptions may fail, and for skewed or multimodal distributions, where mean/SD-based thresholds mislead (Cumming (2014)).

3.5.2 Bootstrap confidence intervals

This study also integrates bootstrap resampling technique (Efron and Tibshirani (1994)), which offers better small-sample validity and interpretability compared to parametric alternatives. Using the affirmation distribution (n=21), 10,000 replacement samples were generated, recalculating percentile ranks each time. The 95% confidence intervals were computed using the 2.5th and 97.5th percentiles of the resulting bootstrap distribution, which yielded accurate error bounds without the need for normalcy assumptions. For example, an observed 98th percentile in jitter variability was reported as 98% (95% CI: 93-100%), offering reliable inference in small-sample speech analysis contexts (Schuller and Batliner (2013)).

3.5.3 Effect size estimation without inference

The magnitudes of the effects were assessed using two valid metrics: (1) standardized differences (the case deviation from the median, scaled by median absolute deviation (MAD)) and (2) distributional overlap (the proportion of affirmations that exceeded the case value).This approach is consistent with current statistical paradigms that place more emphasis on estimation than on testing the null hypothesis (Wasserstein et al. (2019)). Empirical overlap and MAD-based scaling was used to ensure robustness against outliers and small-sample bias in accordance with computational paralinguistics guidelines (Eyben et al. (2015)). These metrics enable reproducible benchmarking for future research involving sparse or unbalanced data.

3.6 Ethical Considerations and Data Privacy

3.6.1 Data ethics and privacy

The Earnings-21 dataset is publicly available. The names of the speakers were pseudonymized. Understanding executive speech patterns prior to credit rating changes can benefit investors, financial analysts, and the overall economy. By employing thorough statistical validation and responsible dissemination to guarantee that results are consistent with moral and societal norms, this study lowers the ethical risks related to the potential misuse of speech data.

3.6.2 FAIR principle and implementation

This study adheres to the FAIR principles by ensuring that all data used are findable, accessible through Github, interoperable via standardized formats and reusable through reproducibility protocols.

3.6.3 Open science practices

This thesis aligns with the principles and infrastructure of the Open Science Framework by adopting transparent, reproducible, and collaborative research practices.

3.6.4 Bias and fairness

To ensure fairness and minimize bias, data is carefully curated to control for confounding variables such as age and gender. Rigorous preprocessing techniques were used to reduce noise and enhance feature extraction accuracy.

3.6.5 Environmental impact

The thesis minimizes environmental impact through efficient computational practices, including lightweight module configurations and selective feature extraction. The overall project pipeline is CPU sufficient, reframing from unnecessarily costing GPU resources.

3.6.6 Reproducibility and replicability

The entire project repository and demonstrator is available on Github.

4 Technical Implementation and System Design

4.1 System Architecture Overview

4.1.1 End-to-end processing pipeline

The pipeline systematically processes earnings calls raw data into acoustic features and statistical parameters from initial preprocessing to parallel acoustic (F0, pause frequency, jitter, etc.) and semantic (FinBERT sentiment) feature extraction and to multimodal validation. Statistical analysis correlates vocal and linguistic patterns with rating outcome using percentile ranking, effect sizes, and case studies, integrating Bootstrap methods to quantify uncertainty.

4.1.2 Component integration design

Preprocessing, feature extraction, and analysis were operated in a formula structure. Acoustic and semantic pipelines operate independently before correlation calculation. While GPU could optimize FinBERT inference, the current implementation relies primarily on CPU processing for acoustic analysis (librosa), statistical computations, and dashboard rendering, as CPU resources sufficiently support the pipeline's core functionality without requiring GPU nodes.

4.1.3 Voice technology framework implementation

The interactive dashboard is live-hosted at streamlit, enabling users to explore acoustic-semantic-credit rating relationships with chosen parameters. Dashboard is frequently use in the financial industry for issuer investor presentations.

4.2 Acoustic Processing Implementation

4.2.1 Audio preprocessing

Wiener filter was chosen for noise reduction given the teleconference recordings with quasi-stationary (e.g., consistent hum, mild office background) background noise and variable speaker-microphone distance, to minimize the mean square error between the estimated and true speech signals (Xia and Bao (2014)). To maintain robust processing across sampling rates and balance time/frequency resolution, the frame size was set at 25 ms, and 50% frame overlap was selected for smooth transition and to minimize artifacts. Hann window reduces spectral leakage. Noise estimation was performed adaptively using both initial silence (first 0.5 seconds) and ongoing voice activity detection. The noise spectrum was computed using Welch's method. A spectral subtraction method computed magnitude spectrum of each frame and subtracted estimated noise spectrum scaled by an oversubtraction factor ($\alpha = 2.0$) and enforced a gain floor of -20 dB. The reconstruction combined cleaned audio using the inverse short-time Fourier transform (ISTFT) with overlap-add. The processed audio was peak-normalized to the range [-0.95, 0.95] to ensure consistent amplitude levels and prevent downstream clipping.

Further, for reproducibility and generalizability, all audios were resampled to 16 kHz, converted to mono channel, and trimmed leading and trailing silence (with 0.1s margin) based on energy threshold.

4.2.2 Feature extraction configuration

Acoustic features (F0_cv, F0_std, pause frequency, jitter_local) were extracted using Praat (via parselmouth), librosa, and OpenSMILE, following standard scientific definitions as implemented in these tools. No custom algorithms or nonstandard formulas were used for better interpretation.

4.2.3 Feature extraction automation

Extraction automation uses parselmouth (Praat) for F0 tracking (75-500Hz range), jitter/shimmer calculation, librosa for temporal and spectral feature extraction, with optional parallel processing of audio segments via Python's multiprocessing module.

4.2.4 Quality control and validation procedures

Quality control is performed via energy thresholding and speech fraction calculation. Automated outlier detection is performed via interquartile range (IQR). The script checks that extracted values fall within physiologically plausible ranges for adult speech, flagging or excluding files with extreme outliers or implausible values (e.g., $F0 > 600$ Hz for adult speech), and cross validate the features extracted by different tools (e.g., Praat vs. openSMILE).

4.3 FinBERT Integration and Sentiment Processing

4.3.1 Transcript preprocessing and segmentation

To ensure reproducibility and semantic precision in sentiment analysis, this study uses token-level transcript files from the `nlp_references` directory of the Earnings-21 dataset. Each file (e.g., `earnings21/transcripts/nlp_references/4320211.nlp`) provides structured annotations per token, including word identity, speaker ID, timestamps, punctuation, case, semantic tags, and WER-related tags. This format enables precise temporal alignment with acoustic features and supports multimodal validation.

For semantic enrichment, the `.nlp` files are aligned with corresponding `.norm.json` and `.wer_tag.json` files (from `transcripts/normalizations/` and `transcripts/wer_tags/`, respectively). The `.norm.json` files provide multiple probabilistic verbalizations (e.g., "twenty twenty" or "two thousand twenty" for "2020") alongside semantic classes (e.g., `YEAR`, `MONEY`), enabling normalization of spoken-language variations into the formal style expected by FinBERT. The `.wer_tag.json` files are used to ensure proper token alignment between `.nlp` and `.norm.json`, mitigating index mismatches and preserving token integrity.

To preserve transcription coherence and temporal consistency, entity-based filtering is not applied to avoid excluding untagged but semantically informative operational language. Instead, the full transcript is reconstructed token-by-token: if a token has a valid normalized verbalization (as indicated in .wer_tag.json and .norm.json), it is used; otherwise, the original token from .nlp is retained. This approach ensures that all tokens are included in the final transcript, maintaining both fidelity to the original speech and compatibility with downstream NLP models. For example, “Fiscal 2020” is rendered as “Fiscal twenty twenty”, while introductory phrases such as “Good morning ladies and gentlemen” are accurately preserved, ensuring context-aware sentiment classification and consistent preprocessing for replicable analysis.

4.3.2 Sentiment score generation and calibration

Sentiment scores were generated with a maximum sequence length of 512 tokens and a sliding window with 50-token overlap to ensure full transcript coverage. For each chunk, softmax probability was computed for negative, positive, and neutral sentiments. Distribution statistics (mean, standard deviation, percentile, and entropy) were calculated.

4.4 Interactive Dashboard Development

The demonstrator for the study is an interactive dashboard which enables users to visually explore the relationship between acoustic stress indicators and the rating actions.

4.5 Statistical Analysis Automation

4.5.1 Bootstrap methodology implementation

Nonparametric bootstrap resampling with 10,000 iterations and a fixed random seed (42) was employed to generate percentile-based confidence intervals for effect size and tie-aware percentile rank estimates. Confidence intervals are computed using NumPy’s np.percentile() function.

4.5.2 Percentile ranking algorithms

Percentile ranks are calculated using the standard definition, accounting for ties. The script iterates over pre-selected acoustic and semantic features, computing summary statistics (mean, median, std, MAD) for both group and baseline distributions. Rank and effect size outputs are stored in structured dictionaries, enabling downstream reporting and visualization.

4.5.3 Robustness and reproducibility

The entire pipeline utilizes fixed random seeds to ensure reproducibility of bootstrap and statistical procedures. Vectorized, library-based algorithms (NumPy, SciPy, pandas) were used to calculate effect size, confidence intervals, and summary statistics. Output files, tables, and visualizations are systematically versioned and saved to support transparent and fully replicable approach. The full codebase is available on Github.

5 Results and Analysis

5.1 Baseline Distribution Characterization

5.1.1 Affirmation cases acoustic baseline establishment

21 affirmed cases' F0 coefficient of variation (F0_cv), F0 standard deviation (F0_std), pause frequency, and jitter local directly extracted by Praat parselmouth and librosa without custom calculation established the acoustic baseline.

Feature	Unit	Extracted by module	Baseline Mean	Baseline Std	Baseline Median	Baseline MAD
F0_cv	(unitless, normalized [0,1])	Librosa, Parselmouth	0.485	0.404	0.500	0.500
F0_std	(unitless, normalized [0,1])	Librosa, Parselmouth	0.470	0.380	0.500	0.468
Pause Frequency	1/sec	Librosa	0.513	0.388	0.530	0.455
Jitter Local	%	Parselmouth	0.448	0.408	0.414	0.414

Table 1: Acoustic features and baseline statistics

The following algorithm are from Praat parselmouth and librosa without custom calculation:

$$\text{F0\_cv} = \frac{\text{std}(F_0)}{\text{mean}(F_0)} = \frac{\sqrt{\frac{\sum_{i=1}^{N}(f_i - \text{mean}(F_0))^2}{N}}}{\frac{\sum_{i=1}^{N} f_i}{N}}$$

where f_i are the pitch (F0) values for each time frame in the call.

$$\text{F0\_std} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_i - \text{mean}(F_0))^2}$$

where f_i are the pitch values.

Proportion of time paused:

$$\text{Pause Frequency} = \frac{\text{Total time spent in pauses}}{\text{Total call duration}}$$

Number of pauses per minute/second:

$$\text{Pause Frequency} = \frac{\text{Number of detected pauses}}{\text{Total call duration}}$$

$$\text{Jitter}_{\text{local}} = \frac{\text{Average absolute difference between consecutive pitch periods}}{\text{Average pitch period}}$$

F0_cv measures the variation of F0 standard deviation against the mean F0. Without F0_cv benchmark in earnings calls so far, Mayew and Venkatachalam (2012) implied that higher F0_cv associate with positive emotions but does not predict negative emotions. Here, the baseline means of 0.485 indicates that, on average, the pitch variability is about 49% of the mean pitch value in affirmed (baseline) calls. The median (0.500) is close to the mean, suggesting a balanced distribution. The substantial standard deviation (0.404) and median absolute deviation (0.500) show moderate spread: some calls are more monotone, others more expressive, but there is no evidence of extreme pitch variability, consistent with controlled professional speech.

F0_std represents the absolute variability of pitch and is normalized to [0,1]. Refer to the literature, Mayew and Venkatachalam (2012) found the regression coefficient for F0_std is positive and highly significant ($p < 0.01$) for positive emotions. Here, the mean of 0.470 and a median of 0.500 indicate moderate pitch fluctuation is consistent across affirmation calls. The standard deviation (0.380) and MAD (0.468) suggest some calls have higher or lower pitch fluctuation, but the overall profile reflect a stable tone without signs of excessive arousal or monotony.

Pause frequency reflects the proportion of time spent in silence or the normalized number of pauses during the call. A mean and median just above 0.5 suggest that, on average, half the time is spent speaking and half in short pauses, or that pauses occur at a regular, moderate rate. The moderate spread (std and MAD) indicates variability among calls but no tendency toward either excessive hesitancy or continuous speech.

Jitter local measures the voice's micro-instability caused by irregular pitch periods. The Praat documentation indicates that the 1.040% threshold for pathology detection set by the Multi-Dimensional Voice Program (MDVP) may be exaggerated because of noise influence. Literature rarely specified benchmark for professional speech anxiety. Here, stable, healthy voices are indicated by a mean of 0.448% and a median of 0.414%. Most calls exhibit minimal jitter, which is consistent with a professional population, while a moderate spread suggests some diversity.

5.1.2 Upgrade/downgrade cases variation analysis

Comparing the two downgrade cases (4346923 and 4384683) and the one upgrade case (4368670) to the baseline, the upgrade case is more expressive and variable in all acoustic features, suggesting positive, confident, or enthusiastic communication. In contrast, the downgrade cases exhibit flatter pitch, less vocal instability, and either frequent or infrequent pausing, which may indicate negative emotion, stress, or caution.

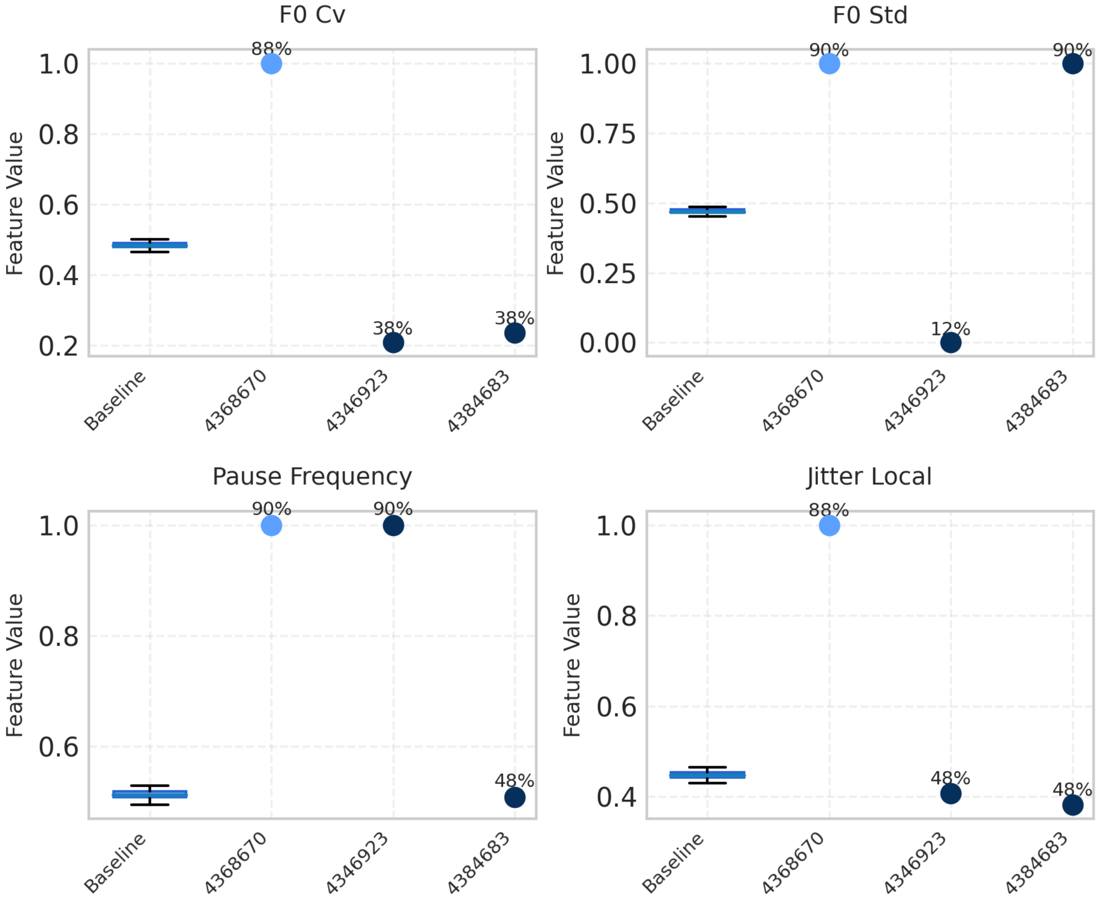


Figure 4: Upgrade/downgrade cases variation analysis

5.1.3 Pairwise acoustic feature correlations

Pairwise acoustic feature correlations are calculated within the baseline group alongside univariate summaries. The results showed moderate to strong associations between some features. For example, speakers who show more stress in their speech also tend to speak less fluently, with more pauses or hesitations.

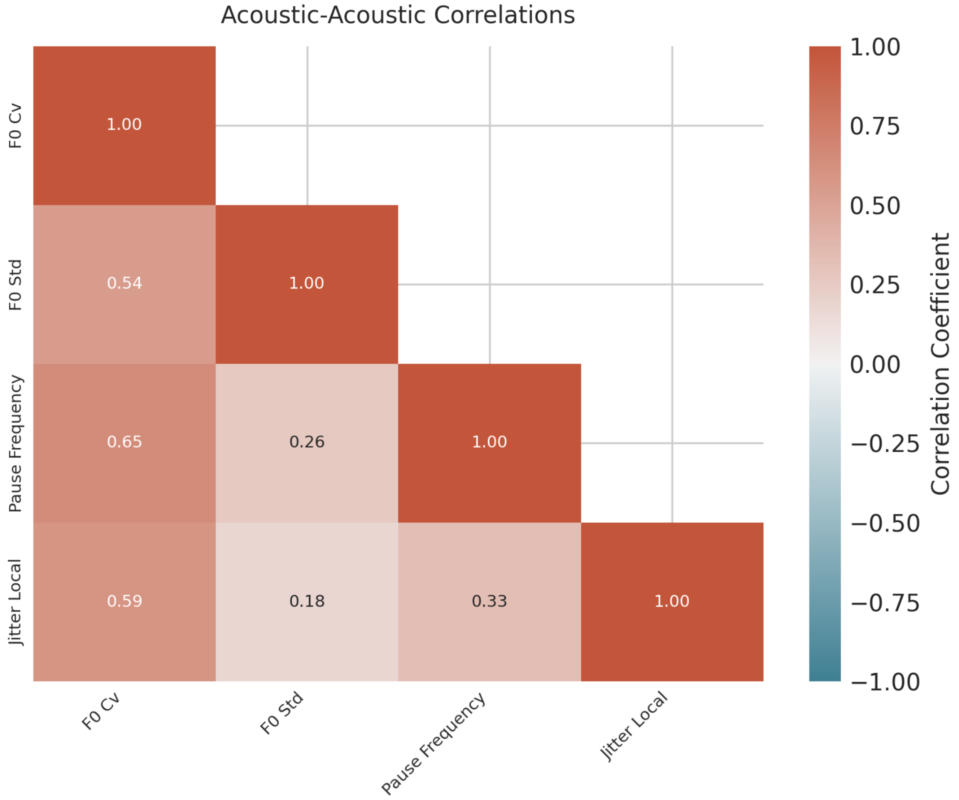


Figure 5: Pairwise acoustic feature correlations

5.1.4 Sector-specific variation analysis

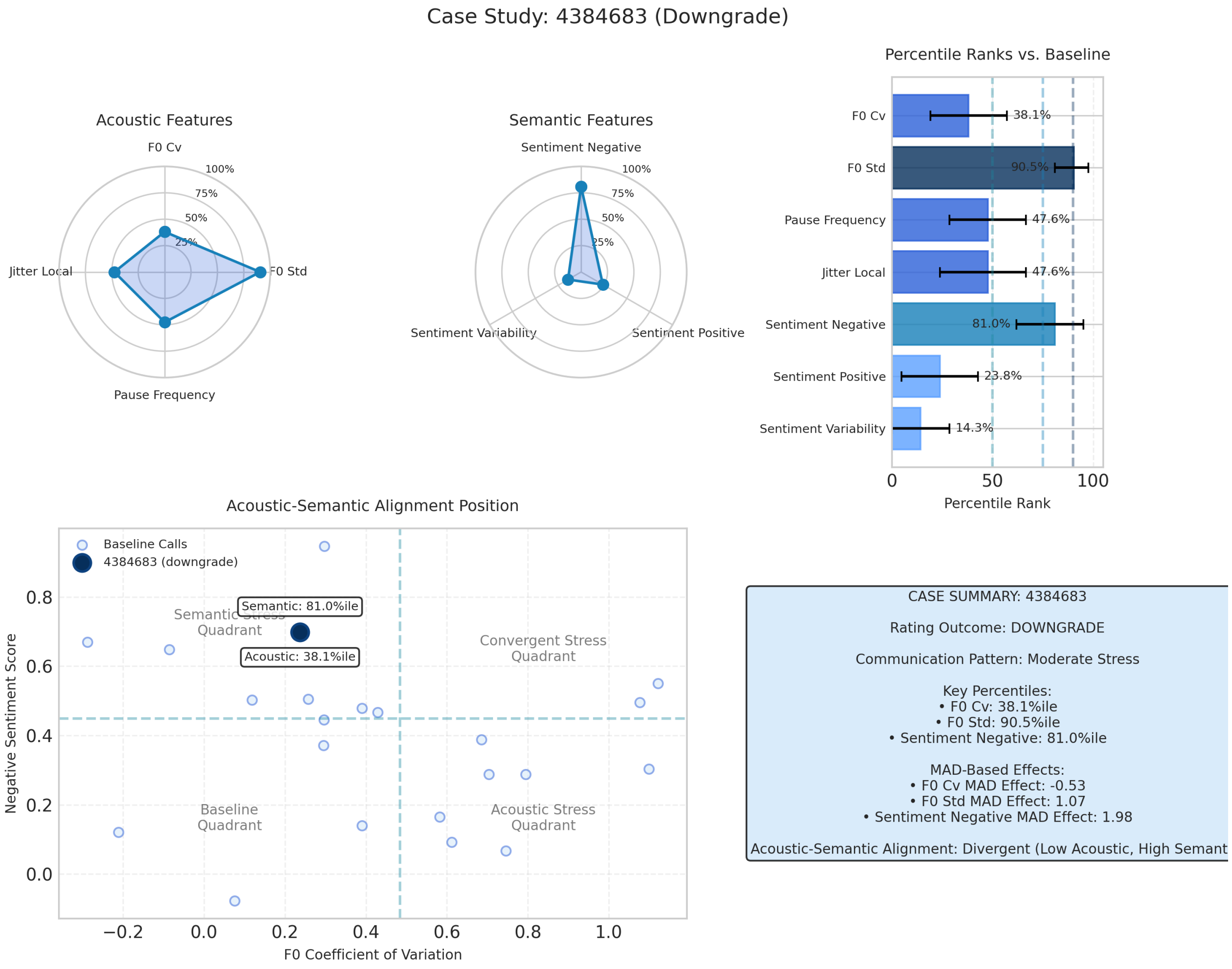
A preliminary sectoral analysis was conducted to look for systematic differences in acoustic baseline distributions, despite the small sample size (3-6 companies) per sector. The results demonstrate that within-sector variability outweighs sectoral effects, suggesting that this sample is appropriate for global baseline characterization. This finding should be interpreted with caution as sector-specific baselines may be supported by larger or more diverse datasets.

5.2 Non-Affirmation Case Studies

5.2.1 Individual downgrade case acoustic profiling

The descriptive investigation of rare non-affirmation cases, specifically downgrades and upgrades, forms the basis of this study. The analysis uses a percentile-based case study methodology instead of any group-level inference because n=1 for upgrades and n=2 for downgrades. For all key features, each non-affirmation case is ranked against the affirmation baseline, and bootstrapped confidence intervals and percentile ranks are calculated.

For example, case 4384683 (downgrade) shows a profile with F0_cv at the 38.1st percentile, F0_std in the 90.5th percentile, and negative sentiment in the 81.0th percentile in relation to affirmations. By looking at each feature in the overall distribution of percentile ranking, this acoustic profile does not sound particularly stressed, but the semantics were more negative than the baseline.



CASE SUMMARY: 4384683

Rating Outcome: DOWNGRADE

Communication Pattern: Moderate Stress

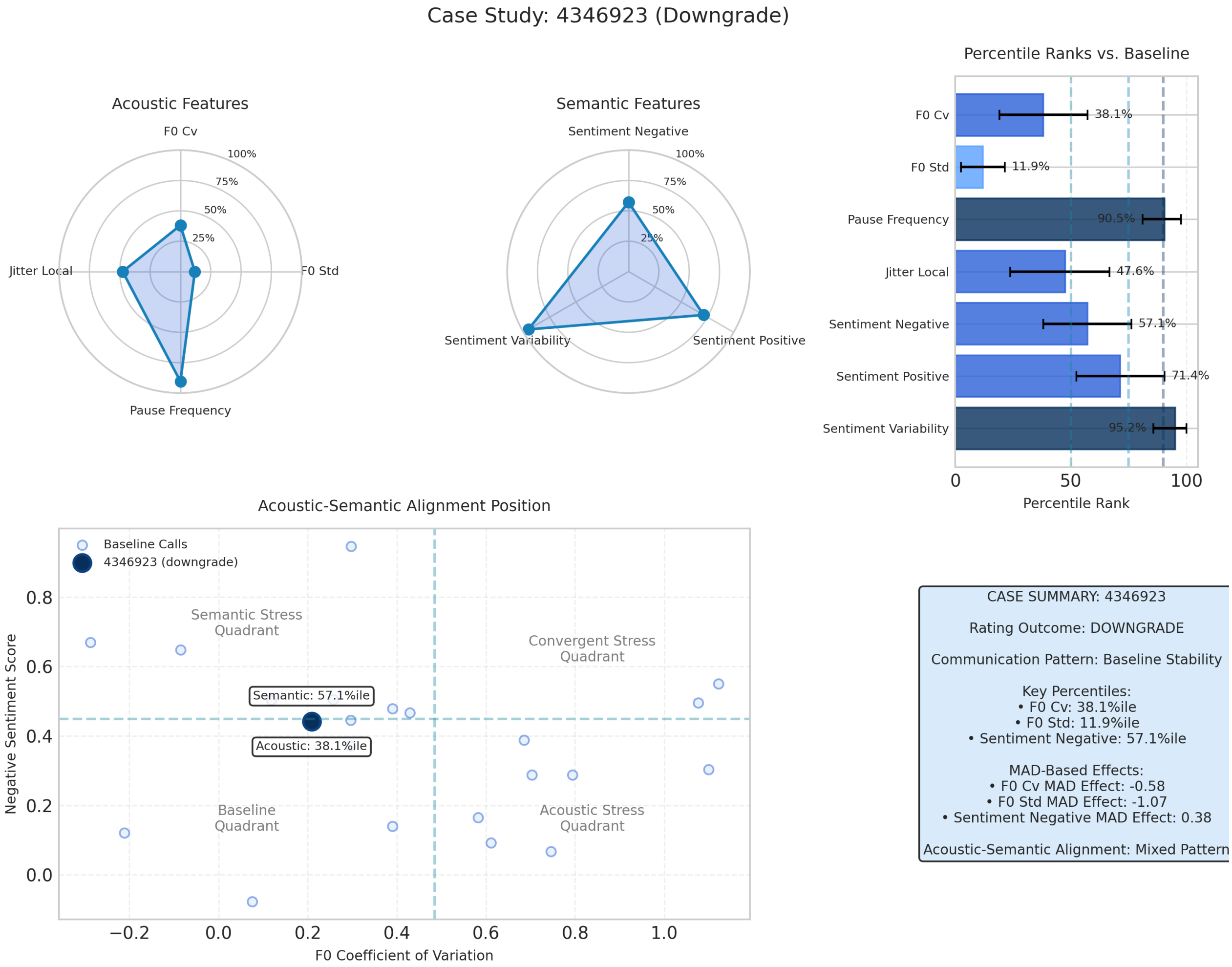
Key Percentiles:
- F0 Cv: 38.1%ile
- F0 Std: 90.5%ile
- Sentiment Negative: 81.0%ile

MAD-Based Effects:
- F0 Cv MAD Effect: -0.53
- F0 Std MAD Effect: 1.07
- Sentiment Negative MAD Effect: 1.98

Acoustic-Semantic Alignment: Divergent (Low Acoustic, High Semant

Figure 6: Downgrade case study I

For 4346923 (downgrade), the F0_cv is at the 38th percentile, and F0_std is exceptionally low at the 12th percentile. However, the pause is more frequent (90th percentile) than the baseline. The semantics are slightly more negative at 57th percentile, but the sentiment variability is extremely high at 95th percentile, suggesting the languages emotions swing between positive and negative. This suggests an overall ambiguous profile, not obviously stressed, but also not entirely calm or negative, and the frequent hesitation indicates uncertainty.



CASE SUMMARY: 4346923

Rating Outcome: DOWNGRADE

Communication Pattern: Baseline Stability

Key Percentiles:
• F0 Cv: 38.1%ile
• F0 Std: 11.9%ile
• Sentiment Negative: 57.1%ile

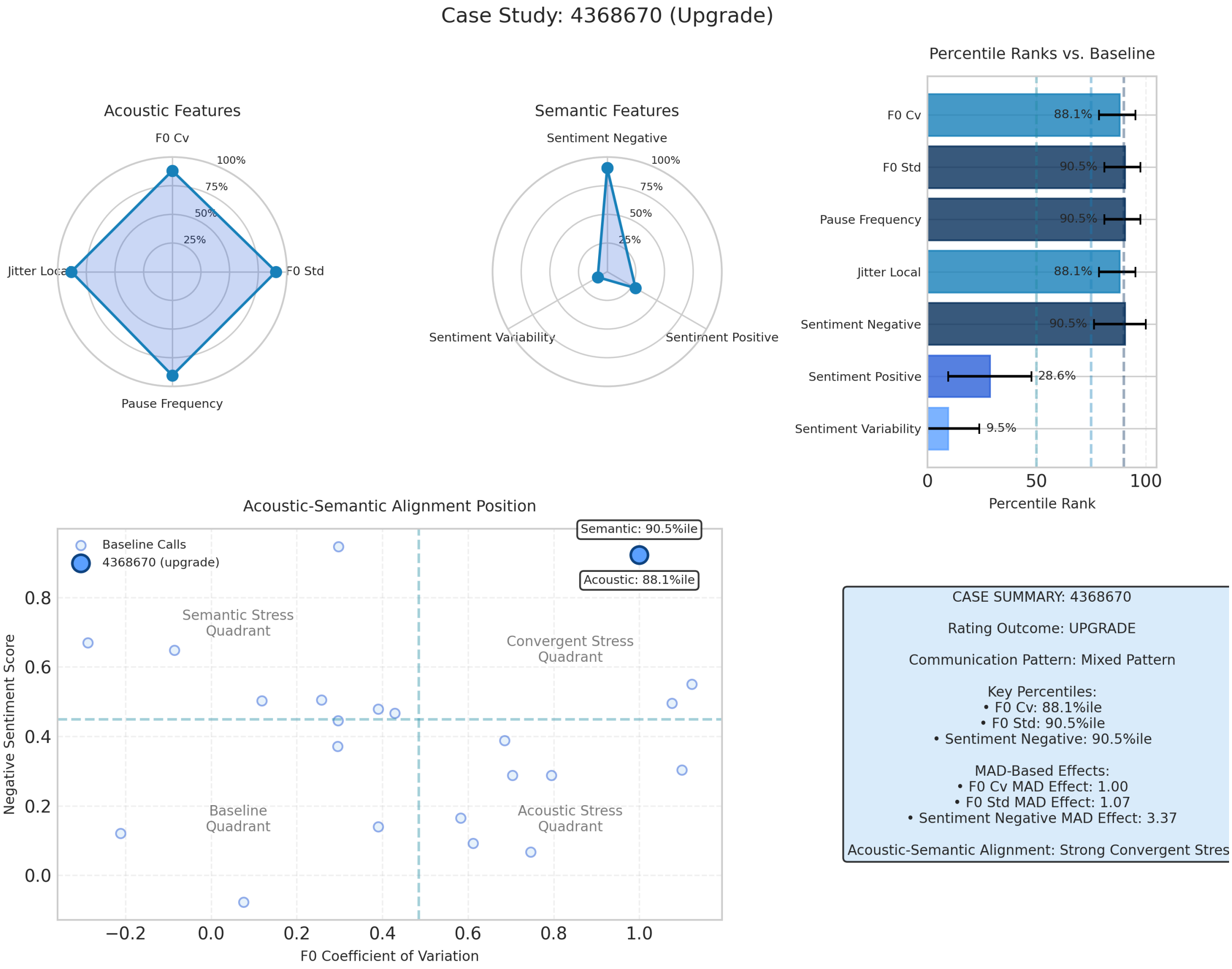
MAD-Based Effects:
• F0 Cv MAD Effect: -0.58
• F0 Std MAD Effect: -1.07
• Sentiment Negative MAD Effect: 0.38

Acoustic-Semantic Alignment: Mixed Pattern

Figure 7: Downgrade case study II

5.2.2 Upgrade case distinctive patterns

For 4368670, the only upgrade case, result is strictly illustrative with n=1. The acoustic stress markers all rank above 88th percentile and the semantics are highly negative (negative sentiment at the 90th percentile). The sentiment variability is low at 10th percentile, and the positive sentiment is low. The communication in this call is overall stressed and negative, although the company received an upgrade. It reveals that tones and languages may send a vastly different signal than the financial outcome may suggest.



CASE SUMMARY: 4368670

Rating Outcome: UPGRADE

Communication Pattern: Mixed Pattern

Key Percentiles:
• F0 Cv: 88.1%ile
• F0 Std: 90.5%ile
• Sentiment Negative: 90.5%ile

MAD-Based Effects:
• F0 Cv MAD Effect: 1.00
• F0 Std MAD Effect: 1.07
• Sentiment Negative MAD Effect: 3.37

Acoustic-Semantic Alignment: Strong Convergent Stres

Figure 8: Upgrade case study

5.3 Acoustic-Semantic Correlation Analysis

5.3.1 FinBERT sentiment validation results

The use of FinBERT sentiment as a directional validator for acoustic stress indicators rather than as a fused or predictive feature is a key methodological attempt of this study. Bivariate correlations between acoustic features and FinBERT-derived negative sentiment are calculated for exploratory purposes only.

The findings show that all cross-modal correlations are weak ($|r| < 0.3$) and non-significant after multiple comparisons are considered. This is consistent with the hypothesis that multi-feature stress patterns might not appear as straightforward linear associations at the call level. Due to the small sample, only the most obvious and one-sided patterns are discernable.

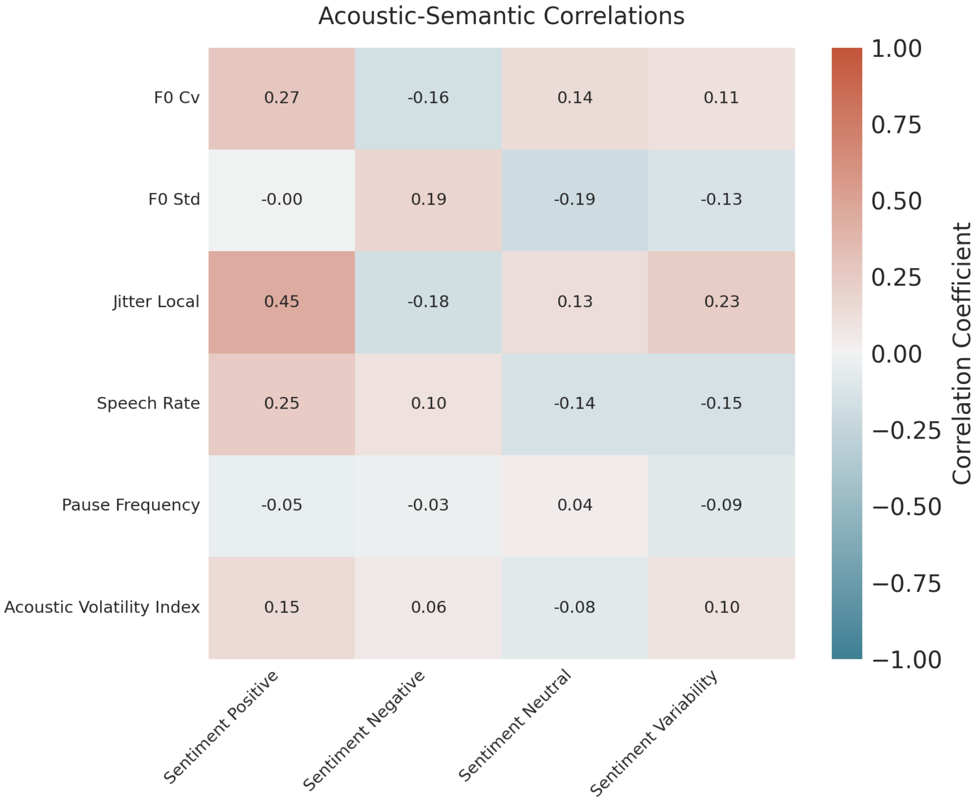


Figure 9: Acoustic-semantic correlation heatmap

5.3.2 Convergent and divergent pattern identification

The case studies reveal that the relationship between what the speakers said and how they said it can match (convergent) or mismatch (divergent). Certain calls sound stressed but do not use negative words, while others use both stressed voice and negative words.

In the following plot, the downgrade cases fall into the divergent or semantic stresses categories, revealing stress in either voice or words, but not both. The upgrade case, on the other hand, is stressed in both and lands in the convergent category.

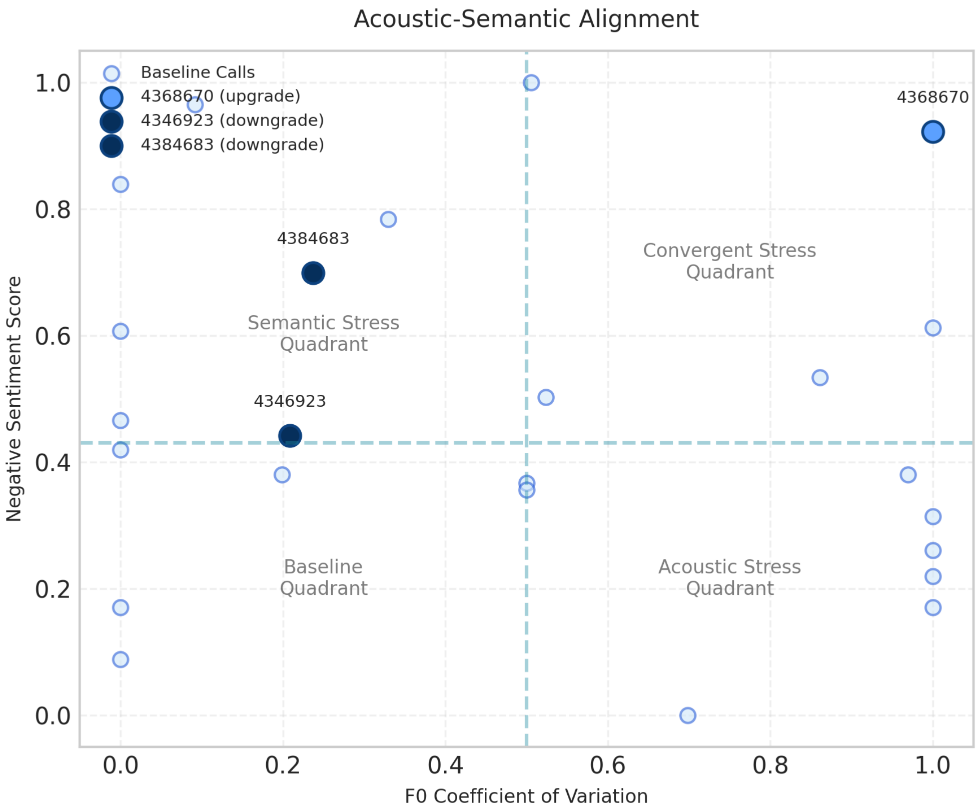


Figure 10: Acoustic-semantic alignment

5.3.3 Acoustic-semantic-credit rating correlations

Most of the relationships are weak or moderate, without any feature with a strong link to the credit ratings. For upgrades, certain acoustic features such as F0_cv and F0_std and pause frequency are higher, indicating more varied speech. For downgrades, there are no significant patterns. For affirmations, no strong correlation is present with some weak link with pause frequency and negative sentiments.

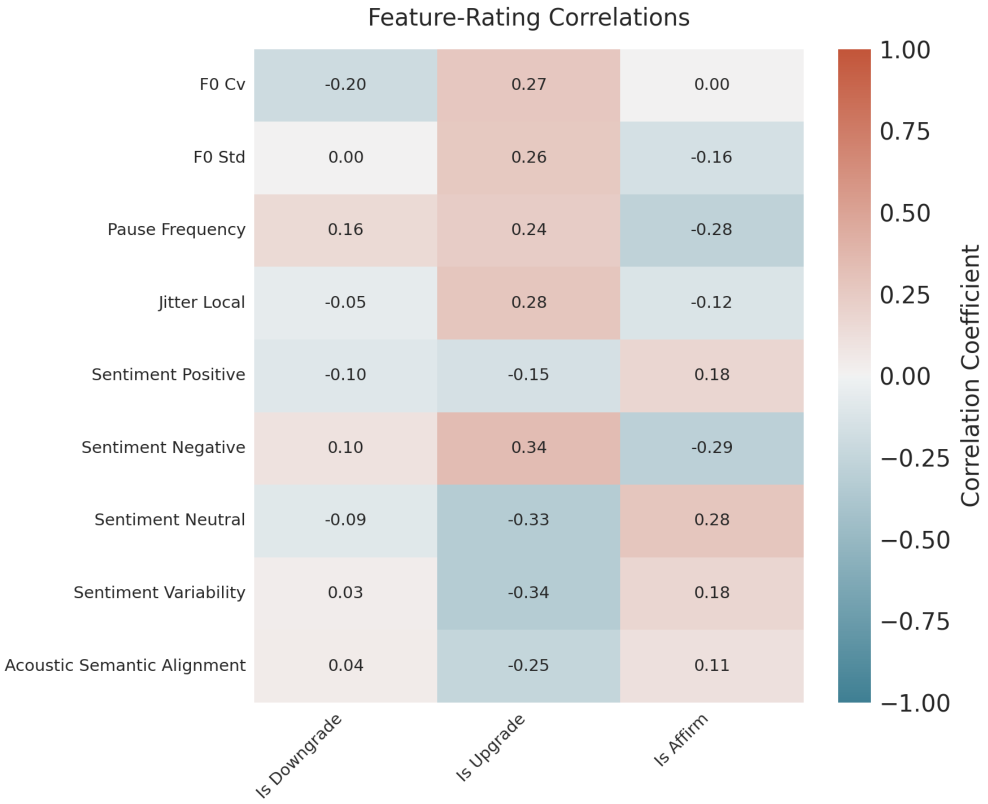


Figure 11: Acoustic-semantic-credit rating correlations

5.3.4 Semantic-semantic correlations

Positive and negative semantics are inversely related, and higher sentiment variability is associated with a greater presence of positive and neutral words in the call (see Appendix B).

6 Discussion

6.1 Interpretation of Acoustic Pattern Findings

6.1.1 Hypothesis validation against descriptive results

The descriptive exploration nature of the acoustic features - credit rating correlation analysis provides partial support for the hypothesis:

H1 predicted that earnings calls preceding credit rating downgrades would present higher F0 variability, pause frequency and jitter compared to the affirmative baseline. The two downgraded cases (4346923 and 4384683) demonstrated heterogeneous patterns which partially denied this hypothesis. 4384683 showed F0 standard deviation at the 90.5th percentile with F0 coefficient of variation only at the 38.1st percentile. This suggests that elevated pitch variability is consistent with stress indicators, while relative pitch variation is below the median affirmation baseline. 4346923 exhibited even more contradictory patterns, with F0 standard deviation at the 12th percentile but the pause frequency at the 90th percentile, indicating stress manifested by hesitation rather than vocal instability.

H2 hypothesized that upgrade cases would show higher acoustic variability coupled with positive sentiments. This is not supported by the single upgrade case (4368670) which demonstrated acoustic variability consistently above the baseline but negative semantics (90th percentile). This suggests that acoustic arousal may not reliably distinguish between positive excitement and negative stress without sentiment validation.

H3's assumption of convergent or divergent acoustic - semantic patterns are partially denied by the findings. The upgrade case exhibited high acoustic arousal coupled with negative sentiment contents - a convergent stress pattern. The downgrade cases presented varied patterns. 4384683 showed moderate acoustic-semantic convergence with both elevated F0 variability and negative sentiment, while 4346923 revealed primarily temporal stress markets (frequent pauses) with modest semantic negativity.

6.1.2 Stress indicator coherence across modalities

A weak cross-modal relationship ($r < 0.3$) is found in the correlation analysis between acoustic features and FinBERT-derived sentiment scores, indicating partial independence between the two modalities. This is consistent with the psychophysiological stress theory, which shows that under complex emotional states, intentional cognitive evaluations (represented in language use) and automatic physiological reactions (observable in vocal features) can diverge. The weak correlations between acoustic and semantic features suggest that financial communication stress can present differently. However, the moderate correlations between acoustic features (e.g., F0 variability showing $r = 0.4$-0.6) show that the acoustic domain is internally consistent.

The validation framework effectively detects divergent patterns, where high acoustic arousal co-occurs with semantically positive or neutral content (indicating controlled anxiety or positive excitement), and convergent patterns, where acoustic and semantic stress align (indicating consistent negative affect). In financial contexts, where physiological arousal can originate from a variety of sources, this multimodal validation approach is useful for interpreting acoustic features.

6.1.3 Individual case study implications

The case-by-case analysis reveals unique communication signatures rather than consistent stress patterns. Calls may exhibit vocal confidence when discussing inconvenient situations that result in positive rating actions, according to the single upgrade case's profile of high acoustic variability with negative semantic content.

The variation seen in downgrade cases suggests that different communication pathways (e.g., vocal instability versus temporal disruption through frequent pausing) may be used to indicate impending negative rating actions. This finding limits the applicability of uniform acoustic stress profiles by indicating that earnings calls' stress responses to worsening business conditions differ. These findings' descriptive character highlights the necessity of larger sample sizes to develop trustworthy frameworks for pattern recognition.

6.2 Methodological Contributions and Validation

6.2.1 Descriptive exploration framework effectiveness

To investigate acoustic-semantic relationships in small-sample financial datasets, this study develops an open and reproducible methodological framework. Without depending on distributional assumptions that are broken in small samples, the percentile ranking method with bootstrap confidence intervals offers reliable statistical characterization. When n=24, the 10,000-iteration bootstrap approach provides more dependable inference than parametric alternatives by effectively quantifying uncertainty around percentile estimates. A significant gap in computational paralinguistics where standardized small-sample protocols are still lacking is filled by the effect size estimation using median absolute deviation scaling, which yields interpretable metrics that can be used for benchmarking in future studies.

6.2.2 Multi-agency rating consensus approach validation

Agency-specific biases that have complicated earlier research are addressed by the consensus-based credit rating classification. The method optimizes the signal-to-noise ratio in the outcome variable by choosing temporally proximate actions and giving rating actions (upgrades/downgrades) precedence over affirmations. In addition to setting baseline expectations for future research, the recording of the temporal intervals (14-606 days) between earnings calls and rating actions offers transparency regarding potential confounding. The methodological decision to concentrate on rated versus unrated companies rather than agency-specific preferences is supported by the validation of consensus ratings against individual agency decisions, which shows that coverage disagreements outweigh directional disagreements. For upcoming research using credit ratings as outcome variables, this methodology offers a reproducible framework.

6.3 Voice Technology Applications and Industry Implications

The integration of acoustic and semantic analysis provides pathways though automatic voice analytics for financial institutions to enhance credit risk monitoring, while such applications must adhere to compliance requirements on biometric data and ethical use and be designed for accurate alignment and robust integration with existing financial IT infrastructure.

6.4 Limitations and Critical Reflections

6.4.1 Sample size and class imbalance

Sample size and the imbalanced distribution is the most significant constrain this study must face. Given limited statistical power, descriptive case study must be interpreted as non-generalizable. Future studies must use larger and more balanced samples to validate any observed pattern and enable robust modelling.

6.4.2 Data coverage limitation

The calls in the dataset were all conducted in 2020. Although including 9 sectors, only 3-6 samples are available in each sector. This means the results do not cover market dynamics or different disclosure cultures. Future research should utilize the larger datasets, such as the SPGISpeech and MAEC, provided credit rating annotation is validated, for generalizable results.

6.4.3 Temporal alignment between calls and ratings

The time gap between calls and subsequent rating actions ranges from 14 to 606 days, complicating the interpretation of correlation. Ideally, selected samples should represent standardized time gap or use modelling to control for temporal uncertainty (e.g., survival analysis if event rates allow).

6.4.4 Speaker and sector heterogeneity

Calls feature 2 to 20 speakers with various roles (e.g., C-suit executives, analysts, operators). Call-level aggregation obscures intra-call, intra-speaker, and role-specific variations. Future work should implement speaker-level and role-level differentiation and use larger datasets for meaningful cross-sector comparisons.

6.4.5 Limited acoustic features and reduced multimodality

Only four acoustic features (e.g., F0_cv, F0_std, pause frequency, jitter local) are used. Full multimodal fusion is avoided given the small data; FinBERT sentiment is only used as a validator. Although justified by sample size, limiting the features and avoiding direct fusion as a methodological choice hinders discovery of richer, potentially non-linear interactions. With larger data, future research should incorporate more comprehensive features and use advanced multimodal learning approaches.

6.4.6 Statistical and methodological constraints

While descriptive statistics, percentile ranking, and bootstrap confidence intervals are used, the methodological approach is appropriate for a case study but cannot provide evidence for causality or predictive validity. However, the study chooses reproducibility and transparency and sets a baseline for future inferential research.

6.4.7 Potential annotation and preprocessing issues

There is potential for annotation errors in aligning transcripts and ratings. No covid-19 related control is implemented, although justified since both speech features and credit ratings reflect contemporaneous company performance and inherently reflect the impact of covid-19. Future studies should invest in more detailed and verified meta-data annotation.

6.4.8 External validity and generalizability

Findings are specified as descriptive and not meant to be generalized, as results are limited to US-listed sector-specific companies’ 2020 earnings calls. The thesis is transparent about its scope. However, the generalizability of the findings is limited pending validation in larger samples.

6.4.9 Technological constrains

Codebase including the demonstrator is dependent on current tools and database formats, with potential for future incompatibility. Although mitigated by open science practices and documentation, future maintenance requires ongoing repository management.

6.4.10 Ethical and privacy considerations

Pseudonymization is performed, but residual risks remain as lack of granular speaker information may limit the ability to control for all ethical risks. GDPR and ethical research conduct is complied, but future studies should monitor evolving standards.

7 Conclusion

7.1 Empirical Method and Findings

This study demonstrates heterogeneous stress patterns across rating actions and explores the methodology of setting empirical baseline for acoustic features in earnings calls correlated with international credit ratings. Using a robust percentile-based methodology, the transparent descriptive exploration framework with bootstrap confidence intervals and FinBERT validation effectively identifies convergent and divergent acoustic-semantic patterns despite the extreme data constraints (n=24), offering reproducible methods for future research in earnings call sentiments' credit rating correlation.

7.2 Future Research Directions

Larger, balanced datasets (200+ events) spanning several years and sectors are needed for future research to support reliable statistical modeling and generalizability. Standardized temporal alignment, advanced multimodal fusion, speaker-level differentiation, comprehensive acoustic features, and improved metadata annotation should all be used in future research. Inferential analysis would be supported, and external validity would be strengthened by longitudinal studies conducted under various market conditions.

References

Alissa, K., & Alzoubi, O. (2022). Financial sentiment analysis based on transformers and majority voting. In *2022 ieee/acs 19th international conference on computer systems and applications (aiccsa)* (pp. 1–4).

Andersson, M., Neves, P., & Nunes, C. (2023). Earnings calls: New evidence on corporate profits, investment and financing conditions. *Economic Bulletin Boxes*.

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint*.

Astuti, R. F., & Alamsyah, A. (2024). Measuring cbdc and defi public discourse using bert and roberta. In *2024 international conference on smart computing, iot and machine learning (siml)* (pp. 150–155).

Baik, B., Kim, A., Kim, D. S., & Yoon, S. (2023). Managers' vocal delivery and Real-Time market reactions in earnings calls. *Chicago Booth Research Paper*, 23–44.

Baik, B., Kim, A. G., Kim, D. S., & Yoon, S. (2024). Vocal delivery quality in earnings conference calls. *Journal of Accounting and Economics*, 101763.

Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech communication*, *46*(3-4), 252–267.

Berteloot, K., Verbeke, W., Castermans, G., Van Gestel, T., Martens, D., & Baesens, B. (2013). A novel credit rating migration modeling approach using macroeconomic indicators. *Journal of Forecasting*, *32*(7), 654–672.

Bornmann, L., Leydesdorff, L., & Mutz, R. (2013). The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits. *Journal of informetrics*, *7*(1), 158–165.

Broś, K. (2023). Acoustic cues to stress perception in spanish–a mismatch negativity study. In *Proc. interspeech 2023* (pp. 2598–2602).

Brown, K., Chen, V. Y., & Kim, M. (2015). Earnings management through real activities choices of firms near the investment-speculative grade borderline. *Journal of Accounting and Public Policy*, *34*(1), 74–94.

Call, A. C., Wang, B., Weng, L., & Wu, Q. (2023). *The listenability of disclosures and firms*.

Cao, Y., Chen, Z., Pei, Q., Lee, N. J., Subbalakshmi, K., & Ndiaye, P. M. (2024). Ecc analyzer: Extract trading signal from earnings conference calls using large language model for stock performance prediction. *arXiv preprint arXiv:2404.18470*.

Cejas, O. A., Azeem, M. I., Abualhaija, S., & Briand, L. C. (2023). Nlp-based automated compliance checking of data processing agreements against gdpr. *IEEE Transactions on Software Engineering*, *49*(9), 4282–4303.

Charlin, V., & Cifuentes, A. (2017). Reliability and agreement of credit ratings in the mexican fixed-income market. *Journal of Credit Risk*, *13*(3).

Chen, Y., Han, D., & Zhou, X. (2023). Mining the emotional information in the audio of earnings conference calls: A deep learning approach for sentiment analysis of securities analysts' follow-up behavior. *International Review of Financial Analysis*, *88*, 102704.

Cohen, J. (2023). Statistical power analysis for the behavioral sciences.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*(1), 7–29.

Dang, V. M. H., & Verma, R. M. (2025). Vocabulary quality in nlp datasets: An autoencoder-based framework across domains and languages. In *International symposium on intelligent data* *analysis* (pp. 288–301).

Das, S., Huang, X., Adeshina, S., Yang, P., & Bachega, L. (2023). Credit risk modeling with graph machine learning. *INFORMS Journal on Data Science*, *2*(2), 197–217.

De Benedicto, S. C., Sugahara, C. R., Silva Filho, C. F., & Sousa, J. E. R. (2018). Organizational communication: a theoretical discussion. *Revista Reuna*, *23*(1), 20–37.

Del Rio, M., Delworth, N., Westerman, R., Huang, M., Bhandari, N., Palakapilly, J., … Jett‘ , M. (2021). Earnings-21: A practical benchmark for asr in the wild. In *Proceedings of interspeech 2021* (pp. 3465–3469). doi: 10.21437/Interspeech.2021-1915

Del Rio, M., Ha, P., Mcnamara, Q., Miller, C., & Chandra, S. (2022). *Earnings-22: A practical benchmark for accents in the wild.*

Doran, J. S., Peterson, D. R., & Price, S. M. (2012). Earnings conference call content and stock price: The case of REITs. *The Journal of Real Estate Finance and Economics*, *45*, 402–434.

Du, K., Xing, F., Mao, R., & Cambria, E. (2024). Financial sentiment analysis: Techniques and applications. *ACM Computing Surveys*, *56*(9), 1–42.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman and Hall/CRC.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., & Truong. (2015). The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, *7*(2), 190–202.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th acm international conference on multimedia* (pp. 1459–1462).

Forbes, E. J., & Pekala, R. J. (1993). Psychophysiological effects of several stress management techniques. *Psychological Reports*, *72*(1), 19–27.

Froot, K., Kang, N., Ozik, G., & Sadka, R. (2017). What do measures of real-time corporate sales say about earnings surprises and post-announcement returns? *Journal of Financial Economics*, *125*(1), 143–162.

Fuller, B. F., Horii, Y., & Conner, D. A. (1992). Validity and reliability of nonverbal voice measures as indicators of stressor‘‘provoked anxiety. *Research in Nursing & Health*, *15*(5), 379–389.

Giddens, C. L., Barron, K. W., Byrd-Craven, J., Clark, K. F., & Winter, A. S. (2013). Vocal indices of stress: a review. *Journal of Voice*, *27*(3), 390–e21.

Gobl, C., & Chasaide, A. N. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech communication*, *40*(1-2), 189–212.

Graham, J. W., & Graham, J. W. (2012). Missing data theory. missing data: Analysis and design. , 3–46.

Grosman, J. S., Furtado, P. H., Rodrigues, A. M., Schardong, G. G., Barbosa, S. D., & Lopes, H. C. (2020). Eras: Improving the quality control in the annotation process for natural language processing tasks. *Information Systems*, *93*, 101553.

Haas, B. D. A. G. . F. R., A. (2022). Stress, hypoglycemia, and the autonomic nervous system. *Autonomic Neuroscience, 240, 102983..*

Haider, F., De La Fuente, S., & Luz, S. (2019). An assessment of paralinguistic acoustic features for detection of alzheimer’s dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing*, *14*(2), 272–281.

Hajek, P., & Munk, M. (2023). Speech emotion recognition and text sentiment analysis for financial distress prediction. *Neural Computing and Applications*, *35*(29), 21463–21477.

Hennig, J. C., Firk, S., & Wolff, M. (2025). Credibility signals from soft information: Evidence from investor reactions to tone in earnings conference calls. *European Accounting Review*, *34*(1), 153–185.

Hirk, R., Hornik, K., & Vana, L. (2019). Multivariate ordinal regression models: an analysis of corporate credit ratings. *Statistical Methods & Applications*, *28*, 507–539.

Hlongwane, R., Ramaboa, K. K., & Mongwe, W. (2024). Enhancing credit scoring accuracy with a comprehensive evaluation of alternative data. *Plos one*, *19*(5).

Hobson, J. L., Mayew, W. J., Peecher, M. E., & Venkatachalam, M. (2017). Improving experienced auditors' detection of deception in ceo narratives. *Journal of Accounting Research*, *55*(5), 1137–1166.

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

Hogg, A. O., Evers, C., Moore, A. H., & Naylor, P. A. (2021). Overlapping speaker segmentation using multiple hypothesis tracking of fundamental frequency. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 1479–1490.

Huang, A. H., Wang, H., & Yang, Y. (2023). Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, *40*(2), 806–841.

Hynes, L., Garvey, J., & O'Brien, F. (n.d.). The anatomy of an earnings call. *Available at SSRN 4696562*.

Ivanitsky, V. P., & Tatyannikov, V. A. (2018). Information asymmetry in financial markets: challenges and threats. *Ekonomika Regiona= Economy of Regions*(4).

Ji, Z., Hou, W., Jin, X., & Li, Z. Y. (2013). Duration weighted gaussian mixture model supervector modeling for robust speaker recognition. In *2013 ninth international conference on natural computation (ICNC)* (pp. 238–241). IEEE.

Johnstone, T., & Scherer, K. R. (2000). Vocal communication of emotion. *Handbook of emotions*, *2*, 220–235.

Kaikaus, J., Hobson, J. L., & Brunner, R. J. (2022). Truth or fiction: Multimodal learning applied to earnings calls. In *Proceedings of the 2022 ieee international conference on big data (big data)* (pp. 3607–3612).

Kappen, M., Van Der Donckt, J., Vanhollebeke, G., Allaert, J., Degraeve, V., Madhu, N., ... Vanderhasselt, M.-A. (2022). Acoustic speech features in social comparison: how stress impacts the way you sound. *Scientific Reports*, *12*(1), 22022.

Kappen, M., Vanhollebeke, G., Van Der Donckt, J., Van Hoecke, S., & Vanderhasselt, M.-A. (2024). Acoustic and prosodic speech features reflect physiological stress but not isolated negative affect: a multi-paradigm study on psychosocial stressors. *Scientific Reports*, *14*(1), 5515.

Kirtac, K., & Germano, G. (2024). Enhanced financial sentiment analysis and trading strategy development using large language models. In *Proceedings of the 14th workshop on computational approaches to subjectivity, sentiment, & social media analysis* (pp. 1–10).

Kisgen, D. J. (2006). Credit ratings and capital structure. *The Journal of Finance*, *61*(3), 1035–1072.

Kumar, Y. S. C. M. A. K. T. . P. P. R., S. (2024). Himal: Multimodal hi erarchical m ulti-task a uxiliary l earning framework for predicting alzheimer's disease progression. *JAMIA open, 7(3), ooae087..*

Lehmann, C., & Tillich, D. (2014). Consensus information and consensus rating: A note on methodological problems of rating aggregation. In *Operations research proceedings 2014: Selected papers of the annual international conference of the german operations research society (GOR)* (pp. 357–362). Germany: Springer International Publishing.

Li, J., Yang, L., Smyth, B., & Dong, R. (2020). Maec: A multimodal aligned earnings conference call dataset for financial risk prediction. In *Proceedings of the 29th acm international conference on information & knowledge management* (pp. 3063–3070).

Li, X., Tao, J., Johnson, M. T., Soltis, J., Savage, A., Leong, K. M., & Newman, J. D. (2007). Stress and emotion classification using jitter and shimmer features. In *2007 ieee international conference on acoustics, speech and signal processing-icassp'07* (Vol. 4, pp. IV–1081).

Liang, P. P., Zadeh, A., & Morency, L. P. (2024). Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, *56*(10), 1–42.

Liao, J., & Shi, H. (2022). Research on joint extraction model of financial product opinion and entities based on roberta. *Electronics*, *11*(9). Retrieved from `https://www.mdpi.com/2079-9292/11/9/1345` doi: 10.3390/electronics11091345

Lipenkova, J. (2022). Choosing the right language model for your nlp use case. *Towards Data Science*.

Lopatta, K., Tchikov, M., & Körner, F. M. (2013). *Misconceptions about credit ratings: An empirical analysis of credit ratings across market sectors and agencies.*

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

Loughran, T., & Mcdonald, B. (2020). Textual analysis in finance. *Annual Review of Financial Economics*, *12*(1), 357–375.

Lu, H., Ehwerhemuepha, L., & Rakovski, C. (2022). A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. *BMC medical research methodology*, *22*(1), 181.

Lu, Q., Du, W., Yang, S., Xu, W., & Zhao, J. L. (2025). Can earnings conference calls tell more lies? a contrastive multimodal dialogue network for advanced financial statement fraud detection. *Decision Support Systems*, *189*, 114381.

Machin, D., Cheung, Y. B., & Parmar, M. (2006). *Survival analysis: a practical approach*. John Wiley & Sons.

Mahon, E., & Lachman, M. E. (2022). Voice biomarkers as indicators of cognitive changes in middle and later adulthood. *Neurobiology of aging*, *119*, 22–35.

Malo, P., Sinha, A., Korhonen, P., & Wallenius, J. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, *65*(4), 782–796.

Maniar, K., Rathod, S., Kumar, A., & Jain, S. K. (2022). A forensic psychological study for detection of deception in financial fraud calls on layered voice analysis (LVATm). *Int J Indian Psychol*, *10*(1), 572–585.

Mathur, P., Goyal, M., Sawhney, R., Mathur, R., Leidner, J. L., Dernoncourt, F., & Manocha, D. (2022). Docfin: Multimodal financial prediction and bias mitigation using semi-structured documents. In *Findings of the association for computational linguistics: Emnlp 2022* (pp. 1933–1940).

Mathur, P., Neerkaje, A., Chhibber, M., Sawhney, R., Guo, F., Dernoncourt, F., & Manocha. (2022). Monopoly: Financial prediction from monetary policy conference videos using multimodal cues. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 2276–2285).

Matthies, A. B. (2013). Empirical research on corporate credit-ratings: A literature review.

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

Miao, J., Lin, J., Luo, T., & Liu, G. (2024). Investor sentiment analysis of financial texts based on gpt and roberta. In *2024 international joint conference on neural networks (ijcnn)* (pp. 1–8).

Miao, M., Wang, Y., Li, J., Jiang, Y., & Yang, Q. (2024). Audio features and crowdfunding success: An empirical study using audio mining. *Journal of Theoretical and Applied Electronic Commerce Research*, *19*(4), 3176–3196.

Mousikou, P., Strycharczuk, P., & Rastle, K. (2024). Acoustic correlates of stress in speech perception. *Journal of Memory and Language*, *136*.

Ng'andu, N. H. (1997). An empirical comparison of statistical tests for assessing the proportional hazards assumption of cox's model. *Statistics in medicine*, *16*(6), 611–626.

Norden, L., & Roscovan, V. (2014). *The dynamics of credit rating disagreement*.

O'Neill, P. K., Lavrukhin, V., Majumdar, S., Noroozi, V., Zhang, Y., Kuchaiev, O., & Kucsko, G. (2021). Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. *arXiv preprint*.

Overes, B. H., & Van Der Wel, M. (2023). Modelling sovereign credit ratings: Evaluating the accuracy and driving factors using machine learning techniques. *Computational Economics*(3), 1273–1303.

Partnoy, F. (2017). What's (still) wrong with credit ratings. *Wash. L. Rev., 92, 1407*.

Patterson, M. G., West, M. A., Shackleton, V. J., Dawson, J. F., Lawthom, R., Maitlis, S., ... Wallace, A. M. (2005). Validating the organizational climate measure: links to managerial practices, productivity and innovation. *Journal of organizational behavior*, *26*(4), 379–408.

Qin, Y., & Yang, Y. (2019). What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 390–401).

Rai, P., Rai, B. S., Pakkala, P. G. R., & Thejaswi, R. A. (2024). Forecasting business status of organizations by analyzing historic earnings call transcripts with the aid of text refinement framework. *Indian Journal of Science and Technology*, *17*(24), 2469–2477.

Raudys, S. J., & Jain, A. K. (1990). Small sample size effects in statistical pattern recognition: recommendations for practitioners and open problems. In *Proceedings. 10th international conference on pattern recognition* (Vol. 1, pp. 417–423). IEEE.

Sangiorgi, F., & Spatt, C. (2017). The economics of credit rating agencies. *Foundations and Trends® in Finance*, *12*(1), 1–116.

Sauter, S., & Jungblut, M. (2023). It's good for our reputation (?!) ‘‘the impact of socio-political ceo communication on corporate reputation. *International Journal of Strategic Communication*. doi: 10.1080/1553118x.2023.2236090

Sawhney, R., Aggarwal, A., Khanna, P., Mathur, P., Jain, T., & Shah, R. R. (2020). Risk forecasting from earnings calls acoustics and network correlations. In *Interspeech* (pp. 2307–2311).

Sawhney, R., Goyal, M., Goel, P., Mathur, P., & Shah, R. (2021). Multimodal multi-speaker merger & acquisition financial modeling: A new task, dataset, and neural baselines. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* (Vol. 1, pp. 6751–6762). Long Papers.

Schuller, B., & Batliner, A. (2013). *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons.

Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F., ... Zhang, Y. (2014). The interspeech 2014 computational paralinguistics challenge: Cognitive & physical load.

Shah, R. S., Chawla, K., Eidnani, D., Shah, A., Du, W., Chava, S., … Yang, D. (2022). When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *arXiv preprint arXiv:2211.00083*.

Shobayo, O., Adeyemi-Longe, S., Popoola, O., & Ogunleye, B. (2024). Innovative sentiment analysis and prediction of stock price using FinBERT, GPT-4 and logistic regression: A Data-Driven approach. *Big Data and Cognitive Computing*, *8*(11).

Shriberg, E. (2001). To ‘‘errrr’is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, *31*(1), 153–169.

Skrlj, B. (2024). From unimodal to multimodal machine learning.

Slapnik, U., & Lončarski, I. (2023). Understanding sovereign credit ratings: Text-based evidence from the credit rating reports. *Journal of International Financial Markets*, *88*.

Throckmorton, C. S., Mayew, W. J., Venkatachalam, M., & Collins, L. M. (2015). Financial fraud detection using vocal, linguistic and financial cues. *Decision Support Systems*, *74*, 78–87.

Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)* (pp. 1–5). Ieee.

Tokdar, S. T., & Kass, R. E. (2010). Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(1), 54–60.

Tomanek, K., & Hahn, U. (2009). Reducing class imbalance during active learning for named entity annotation. In *Proceedings of the fifth international conference on knowledge capture* (pp. 105–112).

Toomet, O., & Henningsen, A. (2008). Sample selection models in r: Package sampleselection. *Journal of statistical software*, *27*, 1–23.

Trouvain, J., & Grice, M. (1999). The effect of tempo on prosodic structure. In *Proc. 14th icphs* (pp. 1067–1070).

Van Puyvelde, M., Neyt, X., Mcglone, F., & Pattyn, N. (2018). Voice stress analysis: A new framework for voice and effort in human performance. *Frontiers in psychology*, *9*.

Verma, P. (2013). Relationship between organizational communication flow and communication climate. *International Journal of Pharmaceutical Sciences and Business Management*, *1*(1), 63–71.

Verma, R. (2024). Building robust ai systems in finance: The indispensable role of data engineering and data quality. *ESP International Journal of Advancements in Computational Technology*, *2*(1), 80–89.

Vlasenko, B., Schuller, B., Wendemuth, A., & Rigoll, G. (2007). Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing. In *Affective computing and intelligent interaction: Second international conference, acii 2007 lisbon, portugal, september 12-14, 2007 proceedings 2* (pp. 139–147).

Wang, J., Zhu, Y., Fan, R., Chu, W., & Alwan, A. (2021). Low resource german ASR with untranscribed data spoken by non-native children. In *INTERSPEECH 2021 shared task SPAPL system.*

Wang, S., Ji, T., He, J., Almutairi, M., Wang, D., Wang, L., & Lu. (2024). *AMA-LSTM: Pioneering robust and fair financial audio analysis for stock volatility prediction.*

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). *Moving to a world beyond ‘‘p¡ 0.05‘‘* (Vol. 73) (No. sup1). Taylor & Francis.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., … Mann, G. (2023).

Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Xia, B., & Bao, C. (2014). Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. *Speech Communication*, *60*, 13–29.

Yang, L., Li, J., Dong, R., Zhang, Y., & Smyth, B. (2022). Numhtml: Numeric-oriented hierarchical transformer model for multi-task financial forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*, 11604–11612.

Yang, Y., Qin, Y., Fan, Y., & Zhang, Z. (2023). Unlocking the power of voice for financial risk prediction: A Theory-Driven deep learning design approach. *Mis Quarterly*(1).

Zhang, S., Li, Y., He, Y., & Liang, R. (2024). Do vocal cues matter in information disclosure? evidence from ipo online roadshows in the sse star market. *International Review of Financial Analysis*, *93*, 103229.

Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y., & Wasinger, R. (2015). Investigation and improvement of multi-layer perceptron neural networks for credit scoring. *Expert Systems with Applications*, *42*(7), 3508–3516.

Zhong, G., Song, H., Wang, R., Sun, L., Liu, D., Pan, J., & Dai. (2022). External text based data augmentation for low-resource speech recognition in the constrained condition of openasr21 challenge. *Proc. Interspeech*, *2022*, 4860–4864.

Appendices

A Earnings Call Datasets Summary

Dataset	Access Type	Modality	Scale	NER	Sentiment/Emotion	Sectors Covered
MAECJ. Li et al. (2020)	Open source	Text + Audio	Large (S&P 1500)	No	No	Broad (1,213 companies)
MDRMQin and Yang (2019)	Not released	Text + Audio	Medium (S&P 500)	No	Vocal (pitch, intensity, etc.)	CEOs (S&P 500 firms)
SER and Text Sentiment for Financial DistressHajek and Munk (2023)	Not released	Text + Audio	Medium (Top 40 US firms)	No	Emotion & Sentiment (FinBERT, NRC)	General sectors (S&P 1500 firms)
SPGISpeechO'Neill et al. (2021)	Open source	Text + Audio	Very Large (5,000 hrs)	No	No	General
Earnings-21Del Rio et al. (2021)	Open source	Text + Audio	39 hrs, 44 calls	Rich (entities: ORG, DATE, TIME, etc.)	No	9 corporate sectors (e.g., Tech, Financial, Healthcare)
Earnings-22Del Rio et al. (2022)	Open source (excluding company metadata)	Text + Audio	119 hrs, 125 files	No	No	Global
ListenabilityCall et al. (2023)	Not released	Text + Audio	56,989 calls	No	ML-based	General
MONOPOLYMathur, Neerkaje, et al. (2022)	Not released	Text + Audio + Video	Medium (6 banks)	Yes	No	Banking
DeepVoiceY. Yang et al. (2023)	Not released	Text + Audio	6,047 calls	No	Vocal cues	S&P 500
AMA-LSTMS. Wang et al. (2024)	Not released	Text + Audio	Medium	Yes	Fairness	General
NumHTMLL. Yang et al. (2022)	Available upon request	Text + Audio	Medium	Numbers	No	General
DocFinMathur, Goyal, et al. (2022)	Not released	Text + Audio + Table	Medium	Yes	Bias Analysis	Broad Financial
M&ASawhney et al. (2021)	Not released	Text + Audio	2016-2020	Yes	No	M&A

Table 2: Overview of datasets used in financial acoustic and textual analysis

B Semantic - semantic correlation

The following semantic-semantic correlation matrix shows that when a call uses more positive words, it usually uses fewer negative words. If the call is more neutral, there is little negative language. When the language in the call has fixed sentiments, it is caused by more positive and neutral words, and fewer negative words. This pattern is in line with expectations about earnings call sentiments.

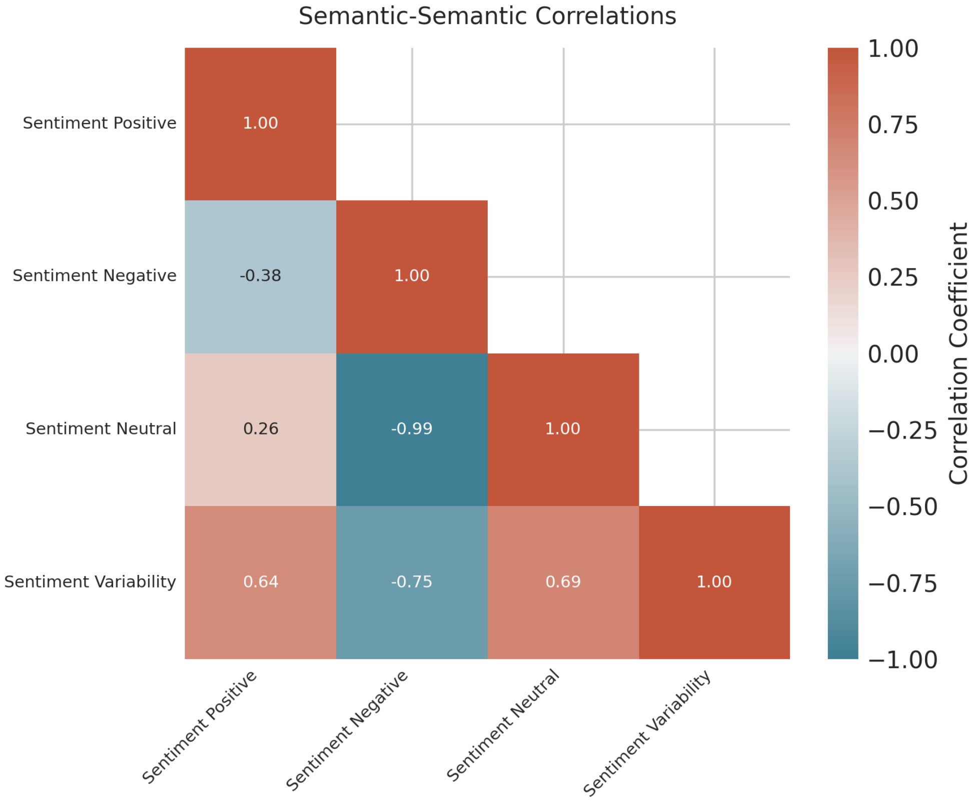


Figure 12: Semantic - semantic correlation

C AI Usage Declaration

I hereby affirm that this Master thesis was composed by myself, that the work herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified, nor has it been published. Where other people's work has been used (from any source: printed, internet or other), this has been carefully acknowledged and referenced.

During the preparation of this thesis, I used GPT-4.1, Claude-Opus-4, and DeepSeek-v3 for the following purposes: consolidate manually-created statistical functions with formatting functions into one Python script for each statistical analysis step, and generate feature extraction and demonstrator codebase for manual editing with Claude-Opus-4; summarize background literature for preliminary review with GPT-4.1; source statistical parameters for domain-specific analysis (e.g., Pearson's r benchmark in financial communication) with DeepSeek-v3. All content was subsequently reviewed, verified, and substantially modified by me.

Tiantian Zhang
11 June 2025