Supplementary Information

Physics-informed modeling of persistent predictive penalty from vocal affect in markets

immediate

September 18, 2025

Contents

List of Tables

List of Figures

1 Main Experimental Results and Statistical Validation

This section provides the detailed empirical evidence supporting the core findings presented in the main text. The following tables offer a granular view of the out-of-sample forecasting performance that underpins the central discovery of a "predictive penalty" associated with vocal affect. We first detail the calculation of our preregistered "uncertainty absorption" metric, followed by multi-horizon performance tables that document the penalty's surprising temporal dynamics. Finally, we present formal statistical tests that validate the significance of these findings.

1.1 Table S1: Absorption Metric Calculation

This table details the components of the primary absorption metric, `Absorb`, a preregistered measure central to testing the paper's "uncertainty-absorption" hypothesis. The metric is designed to quantify the temporal dynamics of vocal affect's predictive power by capturing the degradation of predictive accuracy at intermediate horizons (10–15 days) relative to the short-term (1-day) baseline. The values are calculated as the difference in OOS R^2 between the Multimodal and the **Financial-Only** models, based on data from Table S3.

Supplementary Table S1: Calculation of the Absorption Score. The score captures the sharp decline in predictive power from a slight gain at day 1 to a significant penalty at the 10–15 day mark.

Metric Component	**Value**
ΔR^2 at 1-day horizon (ΔR^2_1)	0.003 625
ΔR^2 at 10-day horizon (ΔR^2_{10})	−0.092 255
ΔR^2 at 15-day horizon (ΔR^2_{15})	−0.091 962
Absorption Score (*Absorb*)	**−0.095 734**

1.2 Table S2: Model Performance Summary (10-Day Horizon)

This table presents key out-of-sample (OOS) performance metrics for the 10-day volatility forecasting horizon. This horizon is particularly relevant as it represents the approximate nadir of the predictive penalty documented in the main paper's Figure 2d. The table compares the full Multimodal model against various baselines and feature-set ablations. A negative OOS R^2 indicates that a model underperforms a simple historical mean forecast, quantitatively defining the "predictive penalty" that is central to our findings.

1.3 Table S3 & S4: Multi-Horizon Performance (Volatility vs. Returns)

The following two tables provide the empirical backbone for a central argument in the paper: vocal affect signals transient, sentiment-driven market risk, not durable information about

Supplementary Table S2: Out-of-sample (OOS) performance for 10-day log realized volatility. The Multimodal model, which incorporates vocal features, shows a substantially greater predictive penalty than the Financial-Only baseline.

Model Configuration	OOS R^2 (log RV, 10d)	95% CI (lower)	95% CI (upper)	OOS MSE (log RV, 10d)	OOS QLIKE (RV, 10d)
Financial-Only	0.006 413	−0.059 574	0.068 039	1.390 651	0.717 551
Financial+Textual	−0.070 656	−0.125 448	−0.019 715	1.498 520	0.820 065
Textual-Only	−0.115 499	−0.157 042	−0.083 979	1.561 283	0.881 203
Acoustic-Only	−0.116 434	−0.158 968	−0.087 306	1.562 592	0.889 370
Acoustic Δ (Q&A − Pres.)	−0.118 504	−0.161 188	−0.087 554	1.565 488	0.900 588
Multimodal	−0.085 842	−0.136 395	−0.039 808	1.519 774	0.831 670

fundamental value. Table S3 documents the full temporal profile of the predictive penalty for volatility, providing the underlying data for Figure 2d in the main text. Table S4 provides the critical counterpart for cumulative abnormal returns (CAR), showing that the same features that degrade volatility forecasts offer no consistent predictive advantage for returns.

Supplementary Table S3: OOS R^2 for predicting log realized volatility across multiple horizons. This table details the temporal signature of the predictive penalty, which emerges after day 1, intensifies significantly by day 10, and persists thereafter.

Model Configuration	OOS R^2 (1d)	OOS R^2 (5d)	OOS R^2 (10d)	OOS R^2 (15d)	OOS R^2 (20d)	OOS R^2 (25d)	OOS R^2 (30d)
Financial-Only	−0.156 768	−0.008 270	0.006 413	0.044 900	0.079 158	0.117 031	0.103 273
Financial+Textual	−0.144 732	−0.051 977	−0.070 656	−0.051 482	−0.033 764	−0.005 579	0.003 306
Textual-Only	−0.175 152	−0.087 566	−0.115 499	−0.119 406	−0.125 143	−0.140 953	−0.151 874
Acoustic-Only	−0.194 556	−0.085 579	−0.116 434	−0.120 716	−0.127 911	−0.143 111	−0.156 975
Acoustic Δ (Q&A − Pres.)	−0.160 960	−0.084 124	−0.118 504	−0.122 372	−0.129 937	−0.171 088	−0.174 539
Multimodal	−0.153 143	−0.054 318	−0.085 842	−0.047 062	−0.048 625	−0.040 662	−0.056 553

Supplementary Table S4: OOS R^2 for predicting Cumulative Abnormal Returns (CAR) across multiple horizons. In stark contrast to the results for volatility, no model configuration shows a consistent predictive advantage for returns, supporting the interpretation that vocal affect conveys risk sentiment rather than fundamental information.

Model Configuration	OOS R^2 (1d)	OOS R^2 (5d)	OOS R^2 (10d)	OOS R^2 (15d)	OOS R^2 (20d)	OOS R^2 (25d)	OOS R^2 (30d)
Financial-Only	−0.124 538	−0.072 626	−0.071 584	−0.141 355	−0.099 290	−0.163 160	−0.162 548
Multimodal	−0.126 301	−0.083 520	−0.077 033	−0.140 092	−0.107 133	−0.119 120	−0.134 142
Financial+Textual	−0.124 337	−0.087 668	−0.080 989	−0.157 075	−0.118 971	−0.139 994	−0.141 870
Acoustic Δ (Q&A − Pres.)	−0.170 999	−0.093 826	−0.084 192	−0.167 789	−0.117 822	−0.130 893	−0.155 353
Textual-Only	−0.177 586	−0.100 352	−0.097 670	−0.184 237	−0.126 989	−0.139 241	−0.167 051
Acoustic-Only	−0.176 464	−0.099 838	−0.110 542	−0.171 163	−0.119 816	−0.133 818	−0.157 717

1.4 Table S5 & S6: Formal Statistical Tests

The following tables move beyond point estimates of OOS R^2 to formally assess the statistical significance of the observed performance differences. Table S5 uses the Diebold-Mariano test for pairwise comparisons against the Financial-Only baseline, while Table S6 uses the Model Confidence Set procedure to identify a set of top-performing models.

Supplementary Table S5: Diebold-Mariano test results comparing models to the Financial-Only baseline. This test assesses the null hypothesis of equal predictive accuracy. Panel B shows that for log realized volatility, the degradation in performance from adding acoustic and textual features is statistically significant at most horizons, formally validating the predictive penalty.

Target	Horizon (days)	Model vs. baseline	Mean loss diff	t-stat	p-value	95% CI (lower)	95% CI (upper)	N
Panel A: Cumulative Abnormal Returns (CAR)								
CAR	1	Acoustic-Only	0.000 133	3.3258	0.0009	0.000 05	0.000 21	1554
		Financial+Textual	−0.000 001	−0.0405	0.9677	−0.000 03	0.000 02	1554
		Multimodal	0.000 005	0.3037	0.7614	−0.000 02	0.000 03	1554
		Textual-Only	0.000 136	3.4034	0.0007	0.000 06	0.000 21	1554
	5	Acoustic-Only	0.000 160	2.3332	0.0198	0.000 03	0.000 29	1554
		Financial+Textual	0.000 088	1.9290	0.0539	0.000 00	0.000 18	1554
		Multimodal	0.000 064	1.1730	0.2410	−0.000 04	0.000 17	1554
		Textual-Only	0.000 163	2.2649	0.0237	0.000 02	0.000 30	1554
	10	Acoustic-Only	0.000 291	2.9108	0.0037	0.000 09	0.000 49	1554
		Financial+Textual	0.000 070	1.0307	0.3028	−0.000 06	0.000 20	1554
		Multimodal	0.000 041	0.5980	0.5499	−0.000 09	0.000 17	1554
		Textual-Only	0.000 195	2.1563	0.0312	0.000 02	0.000 37	1554
	20	Acoustic-Only	0.000 226	1.8776	0.0606	−0.000 01	0.000 46	1554
		Financial+Textual	0.000 217	2.4288	0.0153	0.000 04	0.000 39	1554
		Multimodal	0.000 086	0.9721	0.3312	−0.000 09	0.000 26	1554
		Textual-Only	0.000 305	2.3415	0.0193	0.000 05	0.000 56	1554
	30	Acoustic-Only	−0.000 072	−0.2397	0.8106	−0.000 66	0.000 52	1554
		Financial+Textual	−0.000 309	−1.5909	0.1118	−0.000 69	0.000 07	1554
		Multimodal	−0.000 425	−2.1904	0.0286	−0.000 81	−0.000 04	1554
		Textual-Only	0.000 067	0.2060	0.8368	−0.000 57	0.000 71	1554
Panel B: Log Realized Volatility (`log_RV`)								
log_RV	1	Acoustic-Only	0.260 152	3.6870	0.0002	0.121 75	0.398 55	1554
		Financial+Textual	−0.082 868	−1.7048	0.0884	−0.178 22	0.012 48	1554
		Multimodal	−0.024 962	−0.5498	0.5826	−0.114 02	0.064 10	1554
		Textual-Only	0.126 563	2.5787	0.0100	0.030 29	0.222 84	1554
	5	Acoustic-Only	0.164 154	3.4230	0.0006	0.070 09	0.258 22	1554
		Financial+Textual	0.092 807	3.0919	0.0020	0.033 93	0.151 68	1554
		Multimodal	0.097 776	3.1445	0.0017	0.036 78	0.158 77	1554
		Textual-Only	0.168 373	3.5037	0.0005	0.074 11	0.262 63	1554
	10	Acoustic-Only	0.171 940	3.3385	0.0009	0.070 92	0.272 96	1554
		Financial+Textual	0.107 869	4.1011	0.0000	0.056 28	0.159 46	1554
		Multimodal	0.129 123	4.0949	0.0000	0.067 27	0.190 97	1554
		Textual-Only	0.170 632	3.2191	0.0013	0.066 66	0.274 60	1554
	20	Acoustic-Only	0.228 422	4.4394	0.0000	0.127 50	0.329 35	1554

Supplementary Table S5 – continued from previous page

Target	Horizon (days)	Model vs. Financial-Only	Mean loss diff	t-stat	p-value	95% CI (lower)	95% CI (upper)	N
		Financial+Textual	0.124 567	4.7411	0.0000	0.073 03	0.176 10	1554
		Multimodal	0.140 960	4.3157	0.0000	0.076 89	0.205 03	1554
		Textual-Only	0.225 368	4.3866	0.0000	0.124 59	0.326 14	1554
	30	Acoustic-Only	0.236 027	4.3775	0.0000	0.130 27	0.341 79	1554
		Financial+Textual	0.090 663	2.6880	0.0073	0.024 50	0.156 82	1554
		Multimodal	0.144 951	4.0086	0.0001	0.074 02	0.215 88	1554
		Textual-Only	0.231 400	4.3866	0.0000	0.127 93	0.334 87	1554

Supplementary Table S6: Model Confidence Set for 10-day log RV prediction ($\alpha = 0.05$). The MCS procedure identifies a set of models that are statistically superior to others. The result shows that only a subset of models is included, indicating that other configurations are statistically dominated.

Model in MCS ($\alpha = 0.05$)
Acoustic-Only , Financial+Textual , Financial-Only , Multimodal , Textual-Only

1.5 Table S7, S8 & S9: Sensitivity Analysis of Acoustic Feature Engineering

A core methodological contribution of our work is the robust extraction of nuanced acoustic features using the Physics-Informed Multimodal Model (PIMM). The process of converting raw audio into meaningful quantitative signals—known as feature engineering—is pivotal, yet it often involves numerous discretionary choices that could inadvertently influence the final outcome. Therefore, to bolster the credibility of our conclusions, the following tables detail a comprehensive series of sensitivity analyses. These rigorous tests are specifically designed to ensure that our central finding—the discovery of a persistent predictive penalty linked to vocal affect—is a genuine market phenomenon and not merely an artifact of specific, arbitrary choices made within our feature engineering pipeline.

Specifically, we investigate two primary sources of potential instability. First, we test the model's robustness against variations in sound segmentation methodology, using perturbations in Voice-Activity-Detection (V-A-D) to confirm that the way we define speech segments does not drive the results. Second, we assess the impact of the granularity of the acoustic feature representation by varying the number of acoustic prototypes (K) used for clustering. These tests collectively confirm that our findings are remarkably stable, demonstrating that the predictive penalty is a consistent signal that holds up against different methodological configurations.

Supplementary Table S7: Combined 10-day performance and sensitivity analysis. This table consolidates key robustness checks, showing the stability of the Multimodal model's 10-day performance against perturbations in sound segmentation (V-A-D) and feature representation (acoustic prototypes, K^*).

Model Configuration	**OOS R^2 (10d)**	**V-A-D R^2 (median)**	**V-A-D R^2 (IQR)**	**Prototype K^***	R^2 **@** K^*
Financial-Only	0.006 413	—	—	—	—
Financial+Textual	−0.070 656	—	—	—	—
Textual-Only	−0.115 499	—	—	—	—
Acoustic-Only	−0.116 434	−0.087 418	0.010 864	16	−0.102 554
Acoustic Δ (Q&A − Pres.)	−0.118 504	−0.087 418	0.010 864	16	−0.102 554
Multimodal	−0.085 842	−0.087 418	0.010 864	16	−0.102 554

Supplementary Table S8: Sensitivity of the Multimodal model's OOS R^2 to the number of acoustic prototypes (K). Performance is remarkably stable across a wide range of K values, indicating that the findings are not dependent on a specific level of clustering granularity and that the model is capturing a genuine signal.

K (Number of Prototypes)	**OOS R^2**
16	−0.102 554
32	−0.110 175
64	−0.110 175
128	−0.110 175
256	−0.110 175
512	−0.108 128

Supplementary Table S9: Distribution of the Multimodal model's OOS R^2 under random V-A-D perturbations. The tight distribution of performance metrics under perturbation demonstrates that the model's results are robust to the specific sound segmentation methodology.

Statistic	**Value**
Mean	−0.085 596
Std. Dev.	0.013 406
Min	−0.109 520
25% (Q1)	−0.093 776
50% (Median)	−0.087 419
75% (Q3)	−0.082 924
Max	−0.056 679

1.6 Table S10: Testing an Alternative Mechanism (Emotional Masking)

To explore alternative mechanisms, we investigated whether ”emotional masking”—a conflict between textual and acoustic sentiment—could explain the observed market reaction. This table reports the correlation between changes in our Emotional Masking Index (ΔMI) and future volatility.

Supplementary Table S10: Correlation between the Emotional Masking Index (ΔMI) and future realized volatility. The consistently weak correlation, which is statistically significant across all tested horizons, suggests that emotional masking is not a primary driver of the observed market effects.

Horizon (days)	Correlation (Corr)	CI (lower)	CI (upper)
1	−0.115 173	−0.158 889	−0.067 655
5	−0.094 045	−0.144 617	−0.039 326
10	−0.088 192	−0.147 659	−0.029 919
15	−0.082 008	−0.143 943	−0.022 311
20	−0.077 852	−0.141 411	−0.018 673
25	−0.076 044	−0.139 668	−0.015 786
30	−0.077 762	−0.141 591	−0.017 092

2 Extended Experiments and Causal Probes

This section details extended analyses designed to probe the robustness of our findings and explore potential causal pathways. This includes controlling for a key potential confounder (Question Difficulty) and a preliminary investigation into a causal linkage using an instrumental variable (IV) approach. As noted in the main text, the IV analysis proved inconclusive due to a weak instrument, and we therefore refrain from making causal claims. These experiments correspond to outputs from the script `emovoice_extended_experiments.py`.

2.1 Table S11, S12 & S13: Controlling for Confounders and Probing Causality

A key challenge in this domain is isolating the effect of vocal tone from confounding variables. An analyst's question, for example, could be complex, independently causing both a hesitant vocal response and increased market uncertainty. Tables S11–S13 detail our efforts to address this. We first construct a Question Difficulty Index (QDI) to control for this confound. We then conduct a preliminary instrumental variable (IV) analysis using technical sound glitches as a plausibly exogenous shock to vocal tone.

Supplementary Table S11: Descriptive statistics for the components of the Question Difficulty Index (QDI), constructed to control for the confounding influence of analyst question complexity.

Component	Mean	Std. Dev.	Min	Max
Negative Word Density	0.001 14	0.001 56	0.000 00	0.038 46
Numeric Token Density	0.000 00	0.000 01	0.000 00	0.000 32
Mean Question Length (words)	20.725 10	5.983 18	5.200 00	62.666 67
Type-Token Ratio (TTR)	0.320 84	0.101 07	0.089 63	0.939 39
Question Mark Ratio	0.128 70	0.105 56	0.000 00	1.000 00
Clause Ratio (complexity proxy)	1.279 75	0.435 56	0.000 00	6.000 00

Supplementary Table S12: Summary statistics for the sound glitch instrumental variable (IV). Glitches serve as a plausibly exogenous shock to vocal tone for preliminary causal analysis.

Variable	Mean	Std. Dev.	Min	Max
Glitch rate (text)	0.001 27	0.002 54	0.000 00	0.039 30
Any glitch (indicator)	0.428 97	0.495 07	0.000 00	1.000 00
Glitch bursts (count)	0.456 03	0.555 99	0.000 00	4.000 00
Instrument value	0.001 27	0.002 54	0.000 00	0.039 30

Supplementary Table S13: Two-stage least squares (2SLS) regression results. As discussed in the main text, the weak first-stage F-statistics (well below the conventional threshold of 10) indicate the glitch instrument is not strong enough to support conclusive causal claims. The results are presented for transparency.

Panel A: IV Summary Statistics				
Observations (N)	1440			
F-stat (Δdominance)	2.810 20			
F-stat (Δvalence)	1.496 54			
Stage-2 R^2	0.213 70			
Stage-2 Adj. R^2	0.211 51			
Panel B: Stage-2 Coefficients (Target: log RV 10d)				
Variable	**Coef.**	**Std. Err.**	**t**	**p**
const	−3.147 20	2.510 60	−1.253 57	0.210 00
Pred. Δdominance	247.711 54	177.370 62	1.396 58	0.162 54
Pred. Δvalence	−133.340 80	101.977 62	−1.307 55	0.191 03
Hist. Volatility	9.255 22	4.738 22	1.953 31	0.050 78
QDI	−0.647 10	0.706 66	−0.915 72	0.359 81

2.2 Table S14 & S15: Isolating the Acoustic Signal and Its Economic Magnitude

This section presents two final analyses. Table S14 provides a key robustness check that isolates the unique predictive information in the acoustic channel. Table S15 moves beyond statistical significance to quantify the economic magnitude of the vocal effects.

Supplementary Table S14: Incremental OOS R^2 from adding acoustic features to a Financials+Text baseline. This addresses the alternative explanation that vocal cues are merely a proxy for textual sentiment. The persistence of the "uncertainty absorption" pattern demonstrates that the vocal channel carries a distinct signal related to transient market risk, a key conclusion of our study. Calculated from Table S3 as (Multimodal - Financial+Textual).

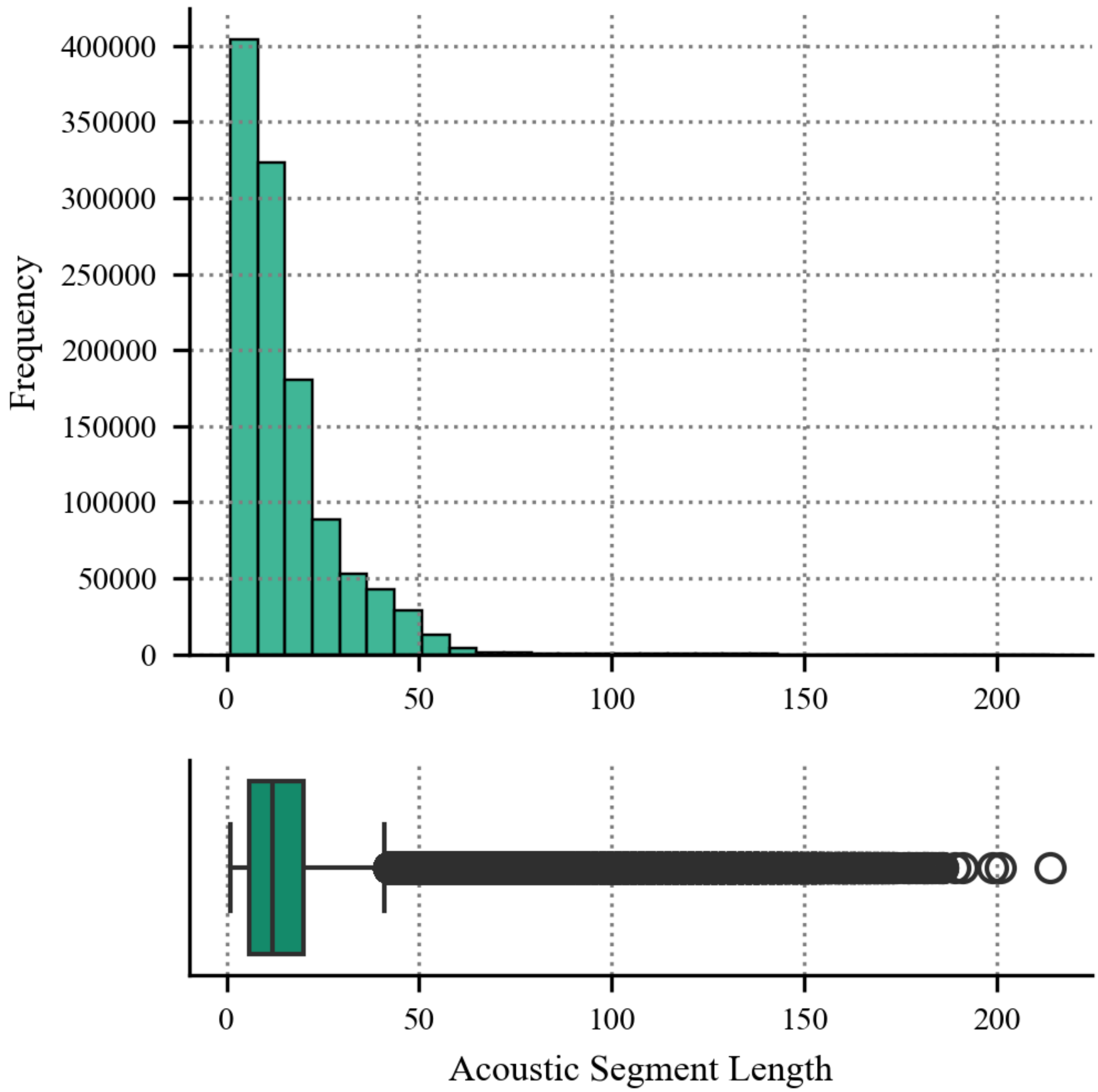
Prediction Horizon (days)	**Incremental ΔR^2 vs. Financials+Text**
1	−0.008 41
5	−0.002 34
10	−0.015 19
15	0.004 42
20	−0.014 86
25	−0.035 08
30	−0.059 86

Supplementary Table S15: Economic magnitude of a one standard deviation decrease in vocal features on 10-day realized volatility. This analysis translates statistical coefficients into economically interpretable units. While these effects quantify a tangible link between vocal tone and market risk, their magnitude is not sufficient to support a profitable trading strategy, underscoring the theoretical nature of our contribution.

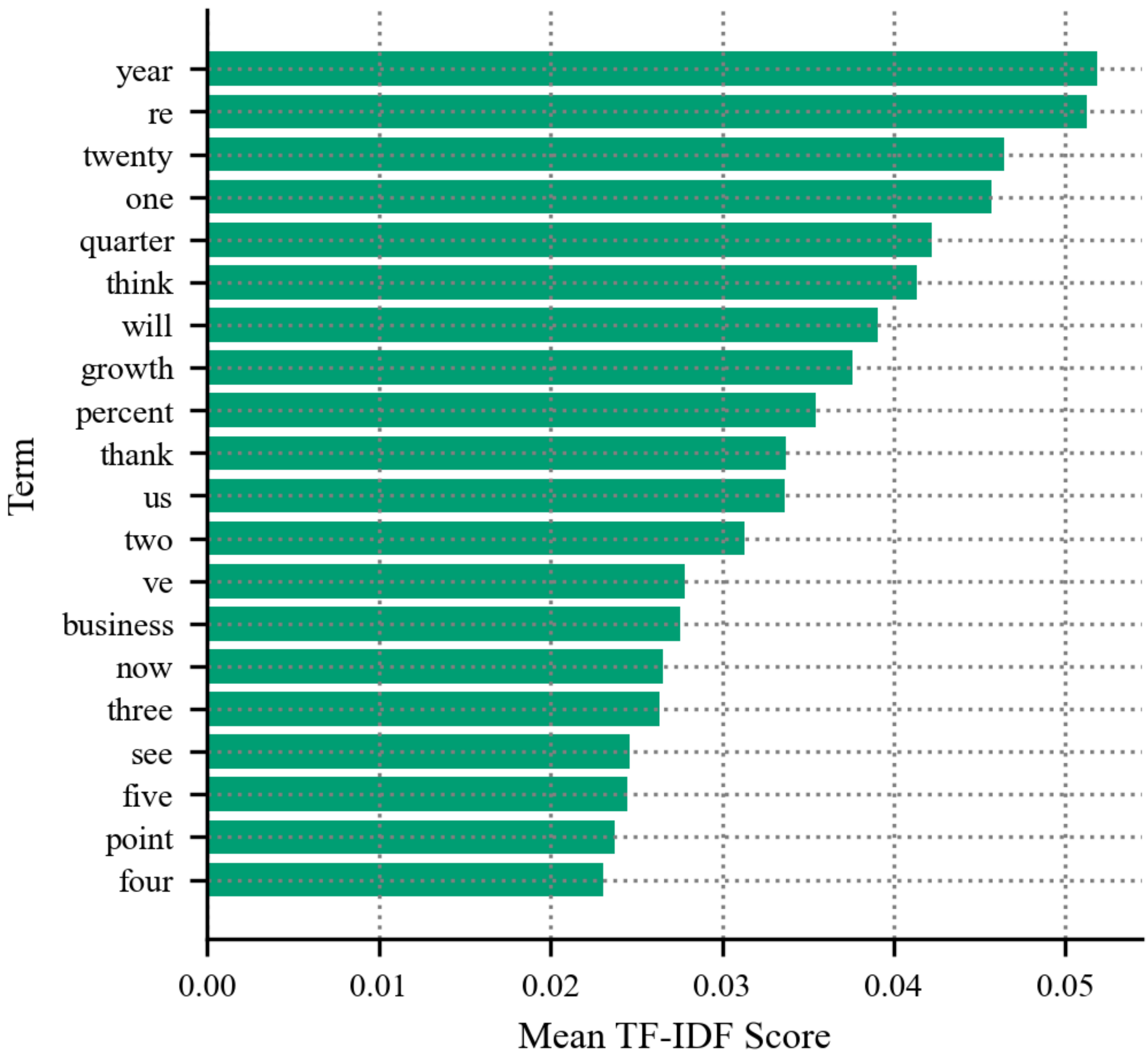
Vocal Feature (1 SD change)	**Δ log(RV)**	**% Change in RV**	**Basis Point Change**
$\downarrow$ in ΔVocal Dominance	0.024 45	2.415 64	241.564 45
$\downarrow$ in ΔVocal Valence	0.045 20	4.419 14	441.913 78

3 Descriptive Statistics Figures

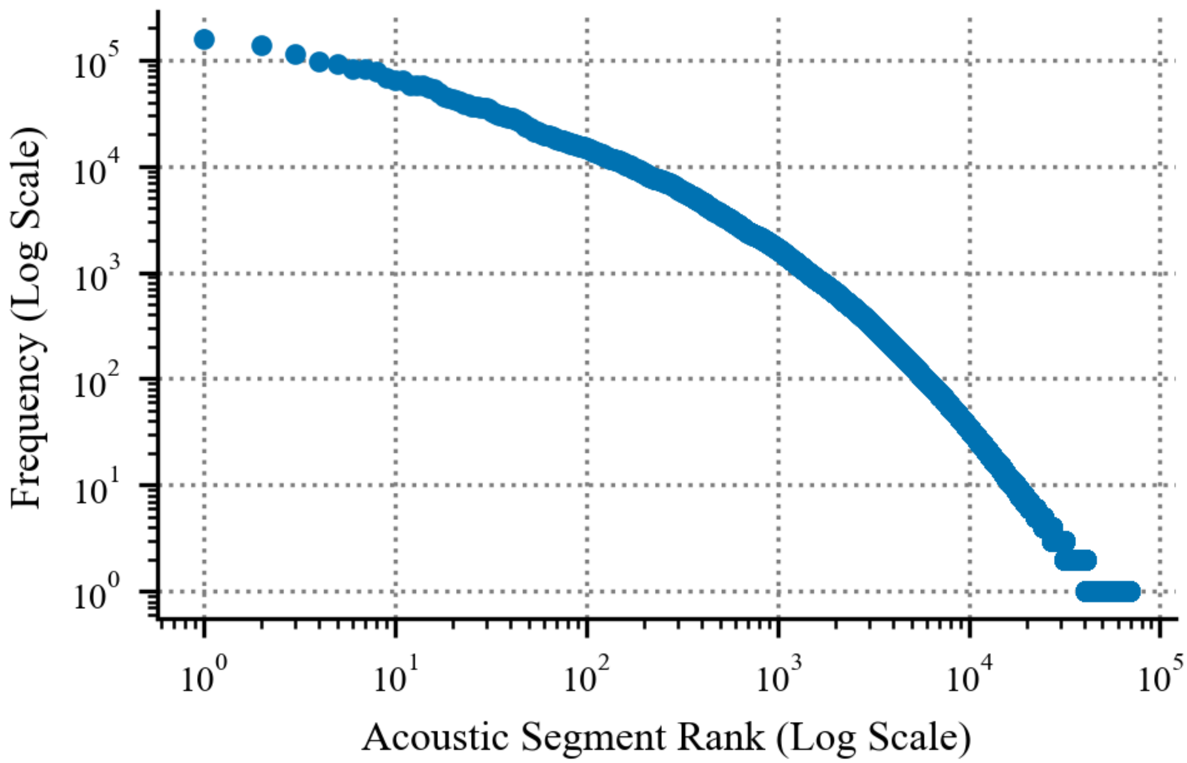
This section provides figures that visualize the fundamental properties of the earnings call dataset. These visualizations offer essential context on the structure of the data, such as the length and frequency of speech segments and the most salient topics of discussion, which are key for understanding the environment in which our models operate. These figures were generated by the script `emovoice_segments_descriptive.py`.



Supplementary Figure S1: Distribution of acoustic segment length in seconds. The high frequency of short utterances reflects the conversational turn-taking that is characteristic of the Q&A sessions.



Supplementary Figure S2: Top 20 most relevant terms from a earnings call transcripts, ranked by TF-IDF scores. This analysis highlights key topics of discussion such as “quarter”, “growth”, and “margin”.



Supplementary Figure S3: Distribution of the number of acoustic segments per earnings call. The plot shows significant variation in the conversational density and length across the calls in our sample.

Supplementary Figure S4: Word cloud of the most frequent terms in the earnings call transcripts. This visualization provides a qualitative summary of the common financial and business terminology used in the calls.