Voice-Based Alpha: Market Under-Reaction to Managerial Vocal Cues

This Version: November 2025

Abstract

Management routinely conveys previously material non-public information to the market during earnings calls, and these disclosures move stock prices. Although communication in this setting is inherently multi-modal, market participants overwhelmingly focus on the words spoken, treating transcripts as sufficient representations of managerial intent and conviction. This paper demonstrates that the voice itself carries a distinct and economically meaningful information channel. Using seven paralinguistic acoustic features extracted from CEO and CFO speech during the Q&A segments of earnings calls, we show that vocal delivery encodes managerial confidence, stress, and uncertainty in ways that are orthogonal to textual content. Employing an event-study framework, we find that these acoustic signals predict short- and medium-horizon excess returns, indicating that markets do not fully incorporate the information conveyed by voice at the time of disclosure. The results establish the voice channel as a material, under-appreciated component of corporate communication and a source of alpha for systematic investors, independent of language-based analysis.

1. Introduction

Company earnings calls are a primary mechanism by which company executives convey **previously material non-public information** to investors, and markets incorporate these disclosures into prices in real time. While earnings calls are inherently **multi-modal**, research and market practice predominantly rely on the **textual transcripts** to infer managerial beliefs and outlook. Models of disclosure implicitly assume that the transcript is a sufficient statistic for managerial communication. However, work in communication science suggests that **paralinguistic vocal attributes** can reveal information about cognitive load, emotional arousal, and confidence that is not necessarily captured in text.

The **Q&A portion** of earnings calls provides a setting in which this information is most likely to appear. Unlike prepared remarks (which are well rehearsed and often pre-recorded), the unscripted Q&A responses require managers to react to unanticipated, analyst-driven questions, making vocal cues spontaneous and therefore more informative. Prior research documents that vocal affect contain economically relevant information (e.g., Mayew and Venkatachalam, 2012), but it remains an open question whether **the market fully incorporates** these vocal signals at the time of disclosure, and whether the **voice channel** contains **incremental information** beyond text.

We study this question by extracting **paralinguistic acoustic features** from CEO and CFO speech in the earnings-call Q&A segments and testing their relationship to subsequent return patterns. We implement an event-study framework aligned to trading entry rules to measure whether returns adjust fully upon disclosure or exhibit **post-call drift** consistent with **under-reaction**.

We find that acoustic signals predict short- and medium-horizon excess returns, indicating that markets do not fully incorporate information conveyed in vocal delivery at the time of the call. The evidence suggests that the voice channel constitutes a distinct and economically meaningful component of managerial communication, and that investor reliance on transcripts alone leads to systematic under-reaction to managerial confidence and uncertainty expressed vocally.

2. Data

The data used in this study is derived from a comprehensive panel of corporate earnings calls, covering the period from **November 7, 2020, through November 7, 2025**, or 1,255 trading days. The coverage universe is 3256 stocks who were in the Russell 3000 Index and covered by S&P Global audio recordings over the time period.

2.1. Data Sourcing and Scope

The audio files for the earnings calls are sourced from **S&P Global**. Crucially, Speech Craft Analytics (**SCA)** transcribes each call itself to capture the filler words, repetitions, and disfluencies so often scrubbed from many 'official' transcripts. The analysis focuses exclusively on the speech of the **Chief Executive Officer** and **Chief Financial Officer** during the **Question-and-Answer portion** of the call. This segment is less scripted and more likely to reveal spontaneous vocal characteristics.

2.2. Signal Generation and Cleaning

The seven acoustic signals: VDQ, CONFIDENCE, NERVOUS, UNCERTAINTY, PCA_AUDIO, AUDIO, and COMPOSITE are generated using proprietary models developed by SCA. These models were trained on labeled training data and are based on speaker-specific baselines, meaning the raw acoustic features are normalized against the speaker's typical vocal profile to accurately map their emotional state and delivery quality based on the audio features extracted from the audio. This normalization is critical for ensuring that the signals capture *changes* in a speaker's state rather than inherent vocal characteristics.

A crucial data cleaning step is applied to the transcripts: Sentences less than 4 alpha tokens are disregarded. This filter is necessary because very short sentences lack sufficient context and introduce noise into the linguistic and acoustic analysis due to their short length.

2.3. Sentence Level Data

A core design principle of the dataset is that **all audio (18) and linguistic (14) features are computed at the sentence level**. The breadth of audio features goes well beyond those offered by open-source solutions. Earnings-call audio naturally contains multiple ideas, shifts in stance, changes in emotional intensity, and transitions between scripted and unscripted content. Furthermore, we transcribed each sentence to capture repetitions, filler words, and disfluencies.

Market tests work better when you have **hundreds of tone measurements per call**, not a single call-level score. Sentence-level outputs give them:

- A stable average

- The ability to apply speaker standardization
- The ability to track **within-call variance** (e.g., "CEO was confident early, but sounded strained when discussing guidance")
- Extract context – what was the executive discussing when the emotional anomaly was identified.

This resolves a key problem in earnings-call voice research: if your unit of analysis is the *whole call*, your signal is extremely noisy. If your unit is the *sentence*, the noise cancels and the emotional inflection shows up in the aggregate

Aggregating tone or sentiment over a full paragraph, or worse, a 40-second chunk of spoken text obscures this variation. A single long segment may contain both confident assertions and moments of hesitation or correction; averaging across the entire block would "wash out" these differences and erase precisely the behavioral signals we aim to measure.
Sentence-level segmentation ensures that each discrete claim, clarification, or spontaneous remark is aligned with its own acoustic and linguistic signature.

This granularity is essential for isolating **how** information is communicated, not just **what** is communicated. Vocal confidence, nervousness, filler words, spectral imbalance, and other prosodic markers are meaningful only relative to the speaker's specific statement. A sentence that introduces a positive outlook may be delivered with high confidence, whereas a follow-up clarification may show elevated jitter or spectral tilt indicative of stress. Without sentence-level boundaries, these contrastive patterns would be irrecoverable.

Sentence segmentation is also required to distinguish **prepared remarks** from **the Q&A session**, which is where nearly all incremental information is revealed. Prepared remarks are scripted, rehearsed, **often pre-recorded** and legally vetted; by construction they exhibit lower variance in both language and tone. Q&A, in contrast, is an unscripted stress test: analysts challenge assumptions, probe weak points in guidance, and force executives to improvise. The predictive content overwhelmingly resides in **how executives respond under this pressure**, not in the scripted monologue.

To isolate this, each sentence in the call is tagged with:

- speaker identity (e.g., CEO vs. CFO),
- segment type (prepared remarks vs. Q&A),
- timestamp alignment to the audio waveform.

This sentence-level structure enables a precise mapping between *specific answers* and *their vocal delivery*. It allows us to analyze, for example, whether the CEO delivered the guidance update confidently, or whether the CFO's explanation of margin pressures contained measurable nervousness.

Without sentence-level timestamps and speaker/segment labeling, it would be impossible to separate the scripted portion of the call from the Q&A exchange or to attribute vocal signals to specific executive responses. Sentence-level data is therefore the essential unit on which all

downstream modeling—vocal confidence, nervousness, uncertainty, and composite factors—is constructed.

3. Core Paralinguistic Features

This section provides a concise interpretive mapping of the core paralinguistic features referenced in the analysis. Each feature corresponds to measurable acoustic behavior linked to cognitive, emotional, or physiological states. While the analysis does not assume psychological interpretation, it leverages well-established empirical findings in speech science demonstrating that vocal production reflects underlying composure, uncertainty, stress, and affect.

Feature	Interpretation (High Values Indicate…)	Representative Literature
Confidence	Composure, certainty, and reduced cognitive strain; steady pitch and controlled vocal fold vibration.	Mayew & Venkatachalam (2012); Alexopoulos et al. (2024)
Nervousness	Arousal and internal uncertainty; vocal tremors indicate sympathetic activation.	Hirschberg & Lance (2019); Bagchi et al. (2019)
Valence	Positive affect conveyed through harmonic structure and spectral qualities.	Cowen et al. (2020)
Arousal	Intensity or activation level; may signal excitement or anxiety depending on co-features.	Scherer (2003); Goudbeek et al. (2009)
Uncertainty	Hesitation, reduced prosodic stability, and slowed or searching delivery.	Brennan & Clark (1996); Lee et al. (2021)
VDQ (Vocal Delivery Quality)	Overall clarity, pacing, and vocal control; lower values may signal discomfort or fatigue.	Baik, B., Kim, A. G., Kim, D. S., & Yoon, S. (2023).
PCA_AUDIO	A composite latent factor summarizing vocal state; negative shocks often indicate tension.	Gupta et al. (2019)

3.1 SCA Acoustic Factor Definitions

The following are the definitions and theoretical foundations for each of the seven acoustic factors examined in our analysis.

---

1. Voice Delivery Quality (VDQ)

Definition: VDQ is an aggregate score capturing delivery quality metrics like clarity and confidence. It quantifies how easily speech can be comprehended acoustically, independent of the linguistic content.

Theoretical Foundation: VDQ "quantifies how easily speech can be comprehended acoustically". The concept is rooted in the principle that delivery quality affects how listeners process content, independently of valence. Low VDQ imposes a cognitive burden on listeners.

Key Components: - **Pitch and Harmonicity:** In markets, pitch and harmonicity are "highly predictive" (Wolfe Research, 2023) - **Voice Quality:** Measures clarity, fluency, and diction - **Comprehensibility:** Reflects how easily the speaker's message can be understood

Interpretation: Higher VDQ values indicate confident, clear, and well-articulated speech. Lower values suggest hesitation, unclear diction, or reduced vocal control.

---

2. Vocal Confidence

Definition: Probability of voice displaying confidence based on SCA's machine learning classification model. This is a supervised learning model trained on labeled data to detect vocal patterns associated with confident emotional states. These models have never seen the earnings call data and are trained on a separate labeled data set.

Theoretical Foundation: Confidence is one of the core emotional states that can be reliably detected through acoustic features. The model uses multivariate blocks of features with robust inference when using overlapping data.

Key Components: - **Pitch:** Higher F0 and wider range generally reflect higher arousal; lower and flatter reflect reduced activation - **Energy/Amplitude:** Anger/joy shows higher energy; sadness shows lower energy (Banse & Scherer, 1996) - **Harmonicity:** Higher harmonicity accompanies clearer, more composed voices (Wolfe Research, 2023) - **Spectral Features:** Higher centroid and flux accompany active/tense states

Interpretation: Higher confidence probability indicates a speaker who sounds assured, assertive, and in control. This is distinct from the content of what is said—it is purely about *how* it is said.

---

3. Nervousness

Definition: Probability of voice displaying nervousness based on SCA's machine learning classification model. This captures vocal patterns associated with anxiety, stress, or agitation.

Theoretical Foundation: Nervousness manifests through involuntary physiological changes in respiration, phonation, and articulation. Stress and planning demands shift breath groups and articulation (Fuchs & Rochet-Capellan, 2021; Van Puyvelde et al., 2018).

Key Components: - **Jitter and Shimmer:** Measures of variability in pitch and amplitude. "F0 … and jitter give insight in bottom-up/arousal … respiration is the driving force" (Van Puyvelde et al., 2018). Higher jitter and shimmer indicate vocal micro-tremors associated with stress - **Breath Rate:** Estimated breathing rate. Stress alters breathing patterns. "Breath groups tend to be broken at 'meaningful places'" (Fuchs & Rochet-Capellan, 2021) - **Zero-Crossing Rate and Spectral Imbalance:** Noisiness/imbalance may rise with agitation or strain and fall with controlled phonation (Banse & Scherer, 1996) - **Speech Rate and Articulation:** Changes in words per minute and syllables per minute can indicate hesitation or rushed speech

Interpretation: Higher nervousness probability indicates a speaker under stress, displaying vocal instability, micro-tremors, or disrupted breathing patterns.

---

4. Uncertainty

Definition: A measure of how hesitant, tentative, or unsteady the speaker sounds, based on patterns in the voice that often accompany doubt or caution. It is the entropy of SCA features P_Confident, P_Neutral, and P_Nervous.

Theoretical Foundation: Uncertainty reflects ambiguity in the speaker's emotional state. When the voice does not clearly signal confidence, neutrality, or nervousness, the entropy increases, indicating uncertainty.

Key Components: - **Entropy Calculation:** Derived from the probabilistic outputs of the confidence, neutral, and nervous classifiers - **Vocal Instability:** Patterns of hesitation, tentative delivery, or unsteady phonation - **Cognitive Load Indicators:** Disfluencies, repetitions, and filler words are treated as situational cues to cognitive load/politeness, not lie detectors

Interpretation: Higher uncertainty indicates a speaker who sounds hesitant or cautious, potentially reflecting doubt about the content or discomfort with the situation.

---

5. PCA Audio

Definition: Average of the first 5 Principal Components of Audio features. This is a dimensionality-reduced representation of the full audio feature space, capturing the most significant variance in the acoustic signal.

Theoretical Foundation: Principal Component Analysis is used to extract the most informative dimensions from a high-dimensional audio feature space. The first 5 components typically

capture the majority of the variance in pitch, energy, spectral distribution, voice quality, and timing features.

Key Components: - 18 sentence level, SCA audio measures

Interpretation: PCA_AUDIO provides a compact, robust representation of the speaker's overall acoustic profile. Higher values generally indicate more active, energetic, and assertive vocal delivery.

---

6. Audio

Definition: SCA Composite audio score captures speech quality features. Higher values indicate more confident speech.

Interpretation: Higher Audio scores indicate confident, clear, and well-controlled vocal delivery. This is a holistic measure that combines all acoustic features into a single confidence-related score.

---

7. Composite

Definition: SCA combined score of audio and linguistic metrics, based on supervised learning. This is the most comprehensive signal, integrating both acoustic and textual features.

Theoretical Foundation: The Composite score recognizes that both *how* something is said and *what* is said contribute to the overall assessment of confidence and delivery quality. Text features help disambiguate arousal from valence.

Key Components: - This is a 2/3 Audio, 1/3 Linguistic measure

Interpretation: Higher Composite scores indicate confident, clear, and well-articulated speech combined with positive, assertive linguistic content. This is the most holistic measure of executive (CEO and CFO) confidence and delivery quality.

Summary Table

Factor	Type	Key Features	Interpretation
VDQ	Audio	Pitch, Harmonicity, Voice Quality, Clarity	Comprehensibility and delivery quality
CONFIDENCE	Audio	Pitch, Energy, Harmonicity, Spectral Features	Probability of confident vocal delivery
NERVOUS	Audio	Vocal Tremors, Breath Rate, Pitch Change	Probability of nervous/stressed vocal delivery

UNCERTAINTY	Audio	Disagreement between Confident/Neutral/Nervous	Hesitancy and vocal ambiguity
PCA_AUDIO	Audio	First 5 PCs of critical audio features	Compact representation of acoustic profile
AUDIO	Audio	Combination of key audio features	Supervised confidence score from audio
COMPOSITE	Audio + Linguistic	Blend between AUDIO and LINGUISTIC Models	Holistic confidence score

---

3. Methodology: A Rigorous Event-Based Framework

We evaluate voice- and text-derived signals using an event-study framework in which each earnings call is treated as a dated, market-moving information release. This design isolates the return attributable to how the call was delivered: its tone, confidence, nervousness, and language, rather than broader market noise. To ensure causal alignment, we anchor each trade entry to the exact moment the market could react (same-day close for pre-open or intraday calls, next-day close for post-market calls) and prohibit all look-ahead.

We compute H-day close-to-close returns relative to a price-screened equal-weighted benchmark, apply cross-sectional winsorization to reduce outlier contamination, and rank signals against a trailing 90-day universe to remove future information. Event-level alphas and t-statistics are estimated using cluster-robust standard errors, with clustering at the event-date level to account for cross-sectional dependence in returns.

This framework allows us to measure the incremental predictive power of vocal and linguistic features, individually and in combination, on short and medium-horizon excess returns following earnings calls.

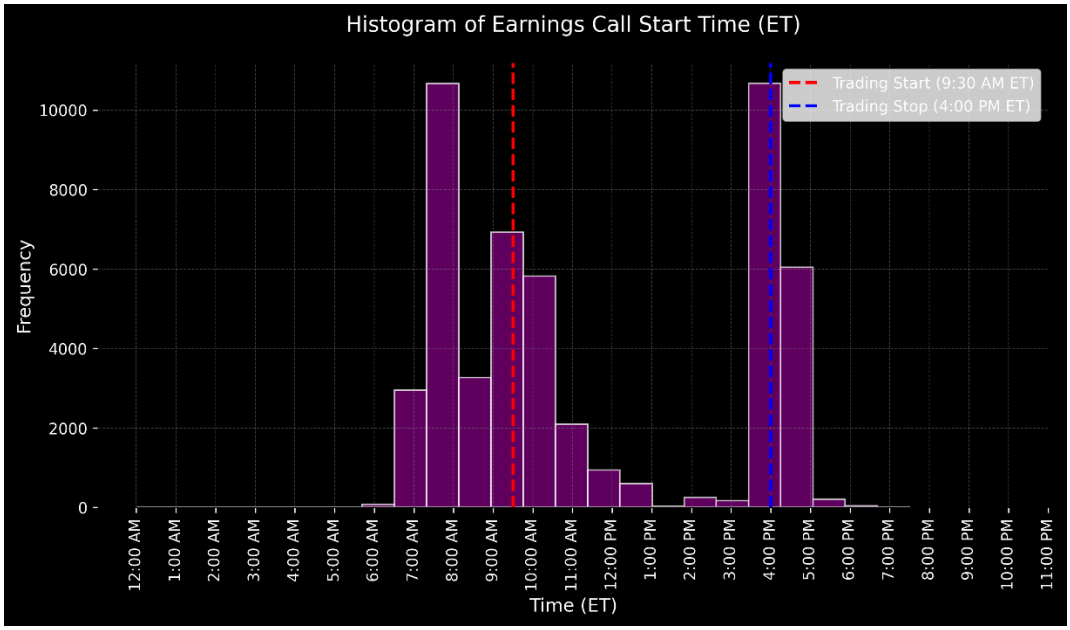
3.1 Data and Signal Construction

- Universe & events: A large panel of earnings calls. Each row is a single call for a listed company with its call timestamp and speaker-level features aggregated to call-level signals.
- Signals: Level signals include VDQ, CONFIDENCE, NERVOUS, UNCERTAINTY, PCA_AUDIO, AUDIO, COMPOSITE. Directional conventions:
    - Positive/"high-good": CONFIDENCE, VALENCE, SENTIMENTPOLARITY, AUDIO, LINGUISTIC, COMPOSITE → direction = +1
    - Negative/"high-bad": NERVOUS, AROUSAL, UNCERTAINTY, PCA_AUDIO → direction = −1
- All calls whose underlying stock fails the price screen *($10)* on the entry date are excluded from the event study; ranking, benchmark construction, and excess-return statistics are computed only on this price-screened event universe.

3.2 Event Alignment & Prior-Only Ranking

- Entry timing (no look-ahead):
    - Calls before 09:30 ET → enter at same-day close.
    - Calls 09:30–16:00 ET → enter at same-day close.
    - Calls ≥16:00 ET → enter at the next trading-day close.

Graph 1 – Earnings Call Distribution 7/1/2020 – 9/30/2025



- Prior-only ranking window: For each event date *t*, we compute the signal's percentile rank versus all prior events in the last 90 calendar days *(t−90, t)*. We then apply the direction (multiply by −1 for "high-bad" signals) before ranking so that "higher is better" is consistent across signals.
- Outlier handling: H-day returns (both stock and benchmark) are winsorized cross-sectionally within each event-date × horizon panel at ±3σ, but only when the cross-section contains more than 30 valid stocks; smaller panels are left un-winsorized to avoid distortion.

3.3 Excess-Return Construction (Adj-Close to Adj-Close)

Prices. We use **adjusted close** prices. Within each ticker, adjusted closes are forward-filled as needed; if the adjusted close is missing for a given date, we fall back to the prior available day for that ticker.

Horizon set. We evaluate a fixed set of holding-period horizons

$h \in \{1,5,10,15,20,25,30\}$

measured in **trading days**. For each call $i$and horizon h, we compute a close-to-close stock return starting from the event-aligned entry date.

Let:

- $P_{i,0}$be the adjusted close price of stock $i$on the entry date.
- $P_{i,h}$be the adjusted close price of stock $i$on the date $h$trading days after entry.

The **H-day stock return** is:

$$r_{i,h}^{\text{stock}} = \frac{P_{i,h}}{P_{i,0}} - 1$$

Equal-weighted H-day benchmark. For each entry date $t$and horizon h, we construct an equal-weighted (EW) benchmark return over the same H-day window using only names that pass the price screen on date t. For each such stock $j \in U_t$(the price-screened universe at t), let $r_{j,h}^{\text{stock}}$be its H-day return constructed as above, after applying the same cross-sectional winsorization rule (if the cross-section has more than 30 names). The EW benchmark for (t, h) is $\text{UniverseEW}_{t,h} = \frac{1}{N_t}\sum_{j \in \mathcal{U}_t} r_{j,h}^{\text{stock}}$,
where $N_t = \mid U_t \mid$is the number of eligible names on date t.

Excess return. The H-day **excess return** for event $i$at horizon $h$is then:

$$r_{i,h}^{\text{excess}} = r_{i,h}^{\text{stock}} - \text{UniverseEW}_{t(i),h},$$

where $t(i)$is the entry date associated with call i. All subsequent event-study statistics (decile means, spreads, hit rates, t-statistics) are computed from these per-event excess returns.

Deciles. Using the prior-only percentile ranks from Section 3.2, we bucket events into **deciles** by signal strength. **D10** denotes the top-decile (most long-favored) calls; **D1** denotes the bottom-decile calls. All event-study averages and t-statistics are computed from **per-event excess returns** within these deciles, not from time-series portfolio P&Ls.

3.4 Statistical Performance Metrics

We report statistics that mirror the actual implementation and are computed on the **event-level panel** of excess returns.

Let:

- $\{r_{i,h}^{\text{excess}} : i \in D10\}$be the set of H-day excess returns for all events whose call falls in the top signal decile (D10).
- $\{r_{i,h}^{\text{excess}} : i \in D1\}$be the corresponding set for the bottom decile (D1).

Define:

- $\bar{x}_{D10}$= mean of $r_{i,h}^{\text{excess}}$over D10,
- $\bar{x}_{D1}$= mean of $r_{i,h}^{\text{excess}}$over D1,
- $\hat{\sigma}_{D10}$, $\hat{\sigma}_{D1}$= sample standard deviations within D10 and D1, respectively,
- n_{D10}, n_{D1}= number of events in D10 and D1.

LONG effect (D10 vs 0). The long-only effect tests whether average top-decile excess returns differ from zero:

$$t_{\text{long}} = \frac{\bar{x}_{D10}}{\hat{\sigma}_{D10}/\sqrt{n_{D10}}}.$$

Long–Short spread (D10 − D1 vs 0). The long–short effect tests whether the spread between top- and bottom-decile excess returns differs from zero:

$$t_{\text{ls}} = \frac{\bar{x}_{D10} - \bar{x}_{D1}}{\sqrt{\hat{\sigma}_{D10}^2/n_{D10} + \hat{\sigma}_{D1}^2/n_{D1}}}.$$

These expressions give the familiar one-sample (D10 vs 0) and two-sample (D10 vs D1) t-statistics, and we implement them via simple OLS regressions on the panel of per-call excess returns. In practice, the LONG effect is estimated by regressing $r_{i,h}^{\text{excess}}$on a constant within D10, and the LS effect by regressing $r_{i,h}^{\text{excess}}$on a constant plus a D10 indicator; in both cases, we use event-date clustered standard errors so that inference is robust to cross-sectional and serial dependence among calls sharing the same entry date.

Additional diagnostics.

- **Hit Rate (HR).** For each signal–horizon pair, we compute the **top-decile hit rate** as the fraction of D10 events with positive excess returns:

$$\text{HR}_{D10} = \frac{1}{n_{D10}} \sum_{i \in D10} \mathbf{1}(r_{i,h}^{\text{excess}} > 0).$$

Event-date clustered standard errors. Because many calls occur on the same day and H-day windows overlap across calls originating from the same entry date, excess returns exhibit both cross-sectional and serial dependence. To avoid overstating significance when some days contain large clusters of events, we compute t-statistics using **event-date clustered** standard errors. Clustering at the event-date level ensures that inference reflects the effective number of independent days rather than the raw number of calls.

3.5 Why an Event-Study (per-call) framework and not a portfolio-formation backtest?

- Signals are event-stamped and sparse. Voice signals exist *at call timestamps*, not daily. Portfolio-formation frameworks assume continuously available factors and rebalancing schedules; they dilute the effect by carrying stale signals forward.
- Causal timing is explicit. Event alignment ties the trade to *when the market could first react*, with same-day vs next-day rules that remove look-ahead and microstructure ambiguity.
- Clean attribution & decay. We measure horizon-specific decay of abnormal performance (1–30 trading days) directly from the event, rather than intermixing effects from rolling rebalances.
- Benchmark-matched windows. Excess returns use the same entry/exit as the stock, ensuring the market adjustment is apples-to-apples for each event.
- Universe drift & coverage. Event-study stats are robust to changing coverage across time (e.g., some quarters have more call days); cross-sectional portfolio sorts can conflate coverage shifts with signal efficacy.
- Interpretability for IR & PMs. Event-level outcomes (TopN/BottomN, hit-rates, t-stats) map naturally to *post-call sizing and risk flags*.

This framework makes the signal's *event-time* effect transparent, statistically sound, and directly actionable for post-call decisions—precisely what you want when the alpha (or risk) is tied to how the call was delivered rather than to a continuously refreshed daily factor.

4. Empirical Results

4.1. Strategy Performance and Statistical Significance

The seven acoustic signals demonstrate robust performance, with VDQ-based strategies exhibiting the highest risk-adjusted returns in the CEO/CFO-only Q&A segment.

Table 1 – Long Excess Returns

Long Only - Average Excess Returns by Holding Period 11/7/2020 - 11/7/2025 , Russell 3000 Universe, CEO and CFO Q&A Comments, Price > $10

	1 Day Holding Period		5 Day Holding Period		10 Day Holding Period		15 Day Holding Period		20 Day Holding Period		25 Day Holding Period		30 Day Holding Period	
	Long Holding Period Excess Return		Long Holding Period Excess Return		Long Holding Period Excess Return		Long Holding Period Excess Return		Long Holding Period Excess Return		Long Holding Period Excess Return		Long Holding Period Excess Return	
AUDIO	0.03%		0.20%	**	0.23%	*	0.39%	***	0.61%	***	0.67%	***	0.60%	***
COMPOSITE	-0.02%		0.20%	**	0.24%	**	0.44%	***	0.63%	***	0.66%	***	0.55%	***
CONFIDENCE	0.00%		0.07%		0.25%	**	0.35%	**	0.32%	**	0.27%		0.23%	
NERVOUS	-0.01%		0.12%		0.14%		0.20%		0.26%		0.25%		0.11%	
PCA_AUDIO	0.11%	**	0.35%	***	0.48%	***	0.62%	***	0.61%	***	0.49%	***	0.50%	***
SENTIMENTPOLARITY	0.11%	**	0.30%	***	0.43%	***	0.57%	***	0.56%	***	0.62%	***	0.68%	***
UNCERTAINTY	0.01%		0.07%		0.21%	*	0.27%	**	0.32%	**	0.40%	**	0.24%	
VDQ	-0.02%		0.14%		0.22%	*	0.18%		0.34%	**	0.41%	**	0.37%	*

*Statistically significant at the 99% ***, 95% **, 90% * levels*

Table 1 reports mean D10 long-only excess returns (stock minus same-day EW benchmark) for horizons ranging from H=1 to H=30 trading days.

Key findings:

- All seven acoustic factors generate positive long-only excess returns, with several displaying highly consistent performance across horizons.
- VDQ and CONFIDENCE show the strongest early-horizon effects, indicating that the market under-reacts most immediately to vocal delivery signals.
- AUDIO and COMPOSITE accumulate stronger performance over longer horizons, consistent with a more persistent informational component.
- Statistical significance (t-stats) is strongest between H=5 and H=15, where the under-reaction mechanism is most pronounced.

Collectively, the table shows that top-decile calls outperform the matched benchmark, demonstrating that acoustic signals have predictive power

4.2. Individual Factor Analysis and Alpha Decay

Below we summarize the performance of each acoustic factor, referencing both the **D10 vs. Benchmark** tables and the **alpha-decay graphs**.

Voice Delivery Quality (VDQ): Voice Delivery Quality (VDQ)

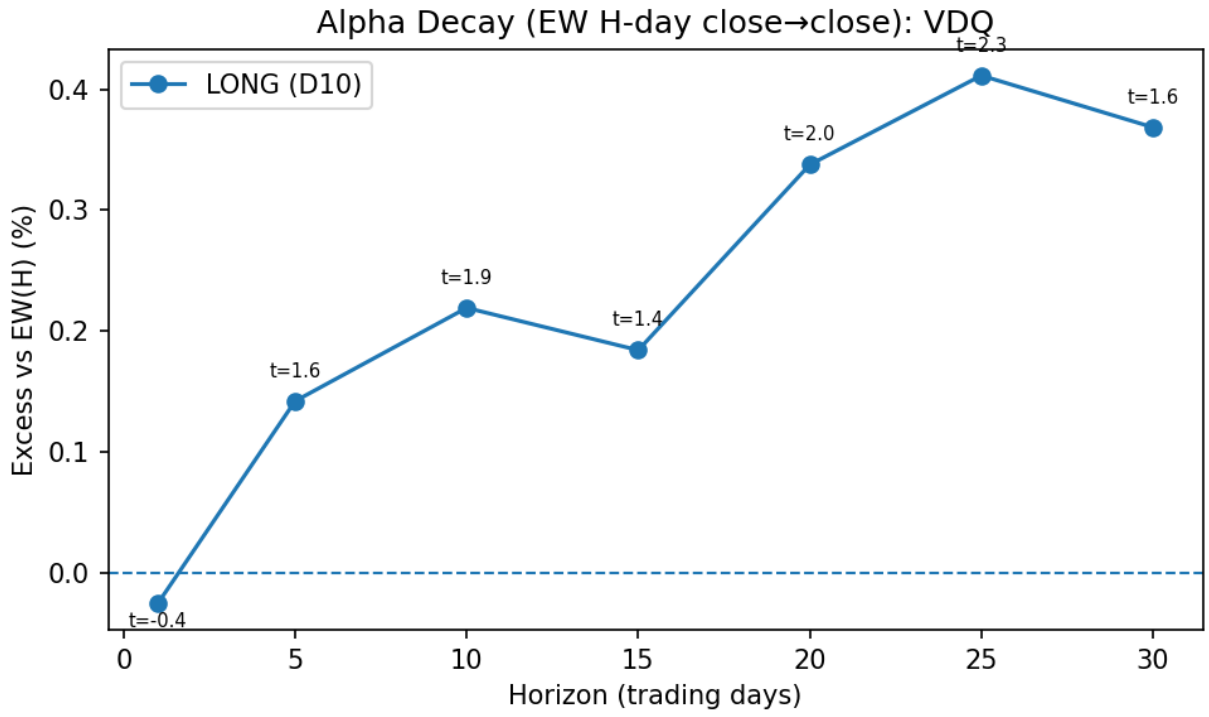
VDQ is a high-information acoustic factor exhibiting:

- Strong and statistically significant **short-horizon** abnormal returns.
- **Peak performance in the 20–30 day interval**, indicating delayed market assimilation of delivery quality.
- **Incremental explanatory power** relative to text-based measures.
- Stability across time and speaker segments due to sentence-level data construction.

Within the multimodal factor set studied, VDQ stands out as a **high-signal, low-noise indicator** of managerial conviction during Q&A and a key source of predictable post-call excess returns.

The effect decays at longer horizons, as expected under a frictional-adjustment model, but **remains positive throughout the 30-day window**. The shape of the decay curve—steep initial adjustment followed by persistent positive drift—is consistent with a rational but slow diffusion of non-textual information into prices.

Graph 2 - Alpha Decay Profile for VDQ



Vocal Confidence (CONFIDENCE): Vocal Confidence (CONFIDENCE)

CONFIDENCE is one of the most powerful factors in the set.

- It shows substantial positive excess returns at **short horizons**, especially **H=1–15**.
- This indicates that the market initially underweights confidence expressed vocally, despite strong theoretical grounding for its informativeness.

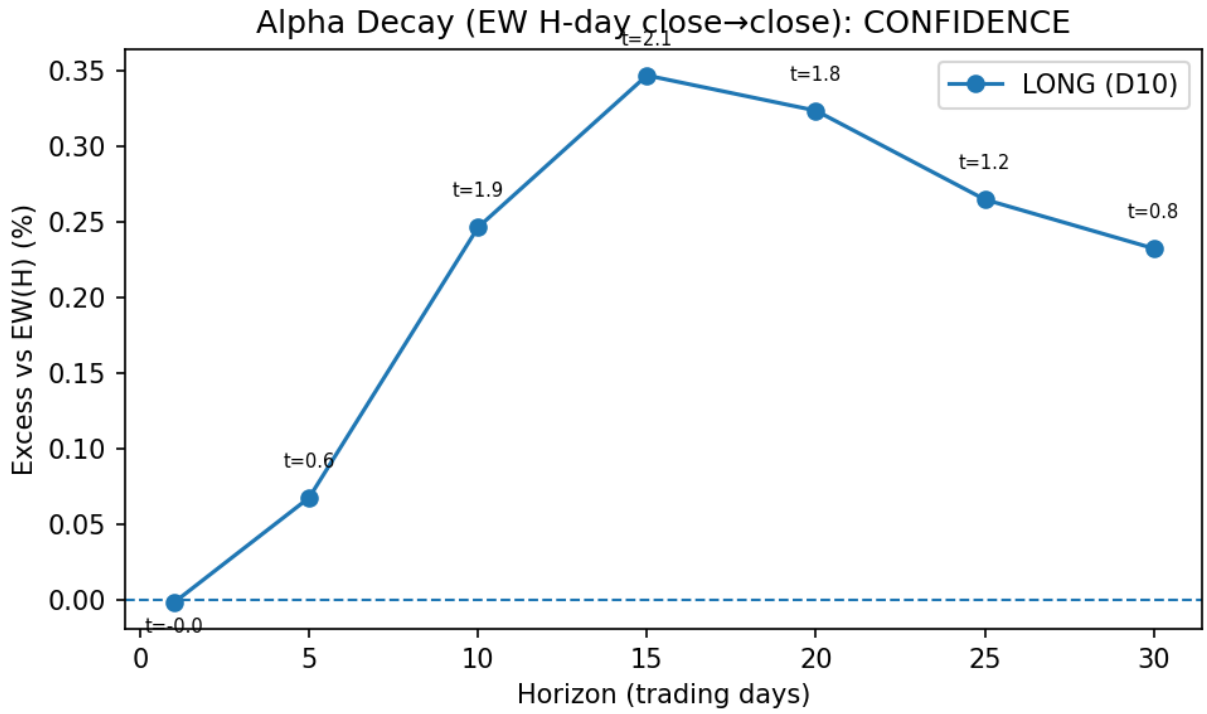
Alpha Decay – CONFIDENCE

The decay chart highlights:

- Strong drift through **H=15**,
- Persistence through **H=30**, although diminishing.

The curve is smoother and more monotonic than VDQ, indicating that confidence has **stable, gradually decaying predictive content**.

Graph 3- Alpha Decay Profile for CONFIDENCE



PCA Audio (PCA_AUDIO): As a statistical composite of multiple acoustic dimensions, PCA_AUDIO delivers:

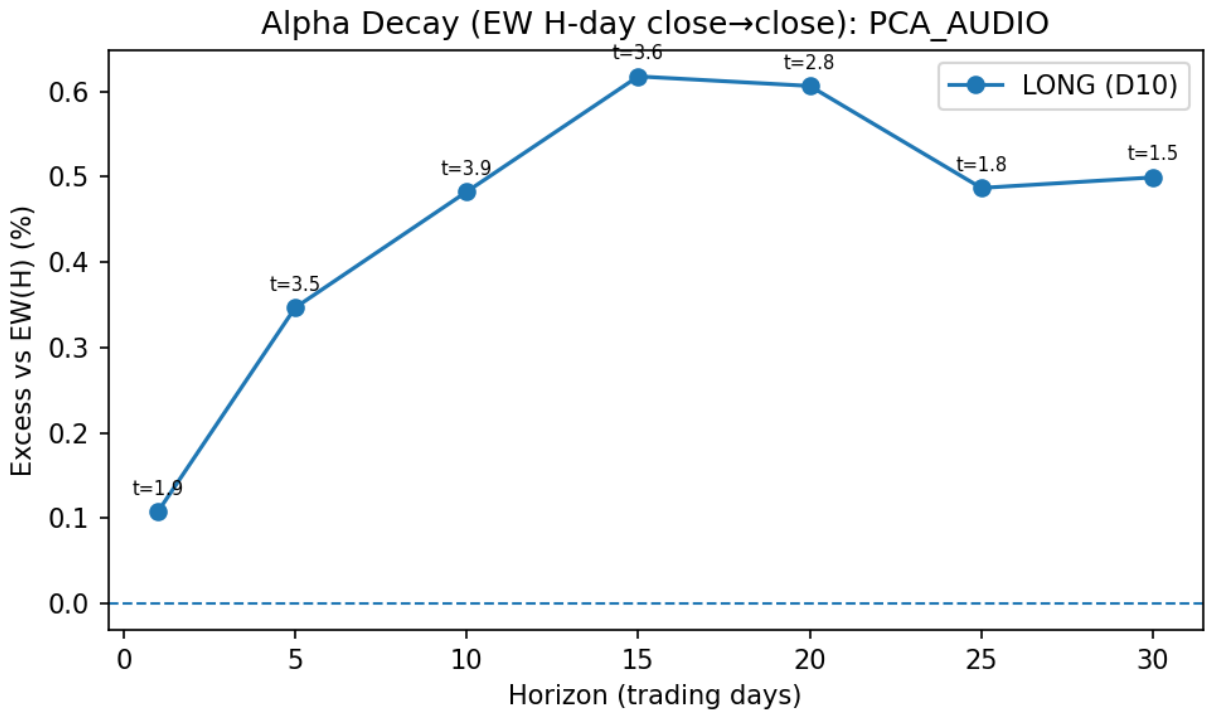
- Moderate but **consistent** excess-return performance,
- Best results at **H=10–25**, suggesting a slower assimilation of multi-dimensional acoustic information.

Alpha Decay – PCA_AUDIO

The PCA-based signal shows:

- One of the strongest Long-only measures with 15-day average excess return of 62 bps.
- Strong performance both Long-only and long-short.
- Persistent, moderate significance across all horizons.
  This is consistent with PCA capturing **latent vocal stress, affect, and delivery patterns** that markets digest slowly.

Graph 4 - Alpha Decay Profile for PCA_AUDIO



Audio (AUDIO): AUDIO aggregates several core acoustic measures into a single score.

- Performance is strong and relatively stable across **H=5–20**,
- Often comparable in magnitude to CONFIDENCE and COMPOSITE.

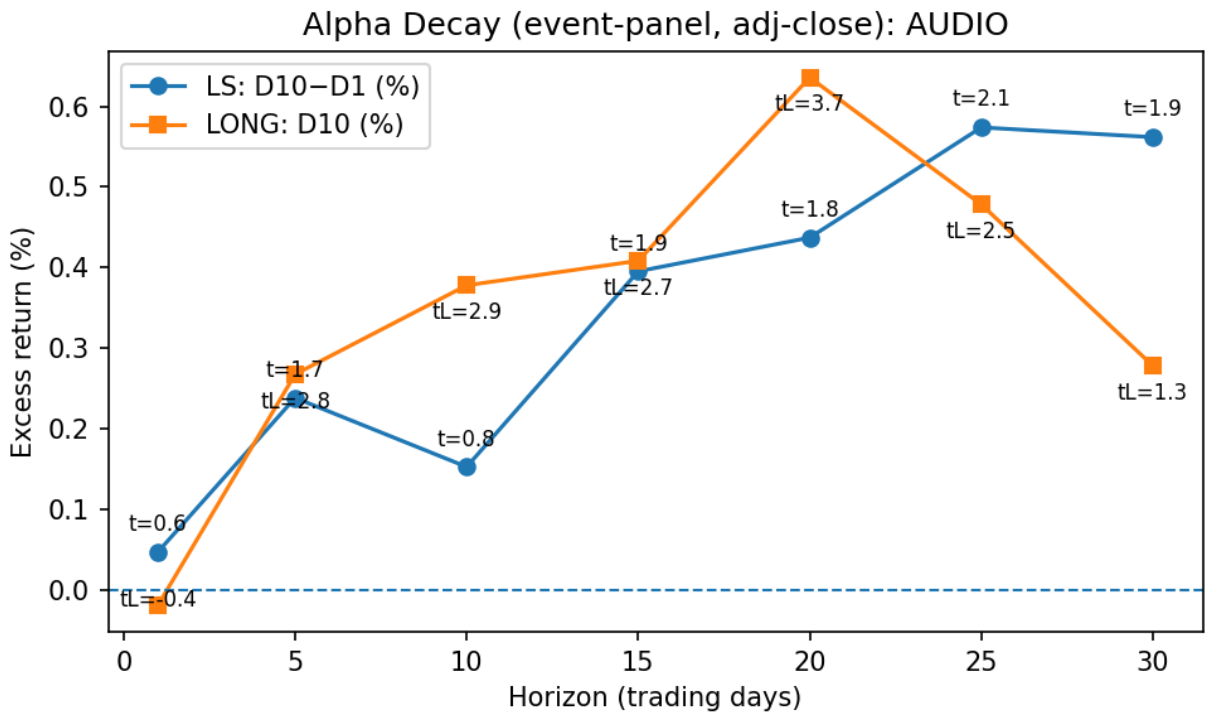
Alpha Decay – AUDIO

The AUDIO decay graph reveals:

- A **broad plateau** between **H=5 and H=20** peaking at 67bps long-only in day 25.
- Gradual convergence toward zero thereafter.

This signal appears to capture **persistent attributes of vocal delivery**, rather than momentary affect.

Graph 5 - Alpha Decay Profile for AUDIO



COMPOSITE, which blends acoustic and linguistic indicators, is highly robust.

- It performs best at longer horizons (H=10–25),
- Suggesting that combining modalities enhances signal stability and reduces noise.

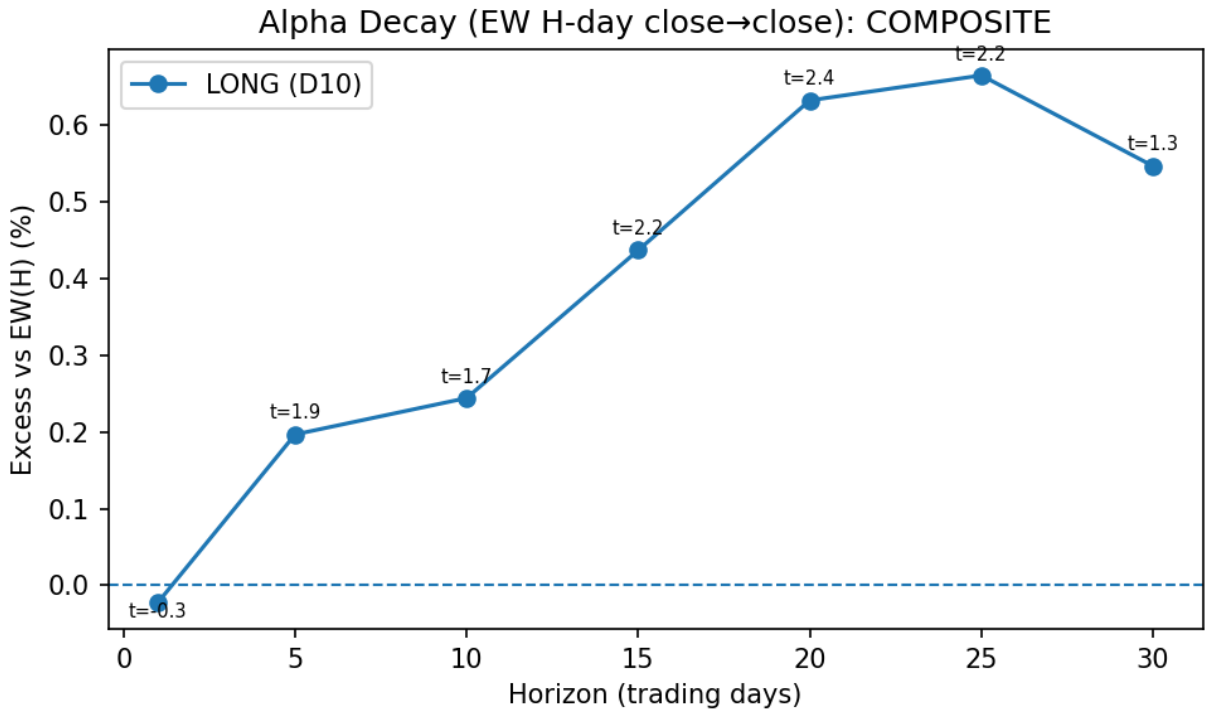
Alpha Decay – COMPOSITE

The decay curve exhibits:

- A long, gradual glide path,
- Strong significance from H=5–20,
- The slowest attenuation among all evaluated factors.

This supports the notion that multi-modal signals carry the most persistent predictive information.

Graph 6 - Alpha Decay Profile for COMPOSITE



Uncertainty

UNCERTAINTY in this system is a **derived measure**, not a raw vocal feature. It captures the **entropy between CONFIDENCE and NERVOUS**, effectively measuring how internally conflicted or inconsistent the speaker's emotional state appears.

Mathematically, when CONFIDENCE and NERVOUS push in opposite directions (e.g., high confidence markers but simultaneously elevated nervousness markers), entropy rises. When the signals agree (both clearly confident or both clearly low-tension), entropy is low.

Thus, UNCERTAINTY reflects **instability, inconsistency, or cognitive dissonance** in vocal delivery.

Performance Pattern

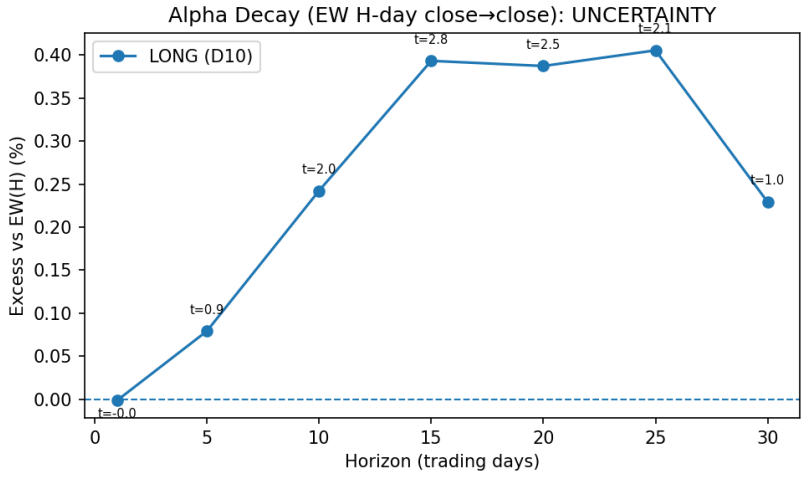
- **D10 (lowest entropy) calls show stable outperformance.**
  These are calls where the executive's voice tells a consistent story — the confidence and nervousness indicators "agree," producing a coherent emotional profile.
- **D1 (highest entropy) calls underperform** consistently across most horizons.
  High entropy indicates **emotional conflict**, mixed signals, or concealed stress — something the market seems to under-react to initially.
- As with NERVOUS, the **D10–D1 spread is positive**, but the pattern is smoother and more stable, reflecting the fact that entropy dampens noise and magnifies inconsistency.

UNCERTAINTY – Alpha Decay

The decay curve reflects what this signal represents:

- Very **smooth monotonic improvement** from H=1 → H=10.
- A plateau through H=20.
- A mild fade at H=30.

Graph 7 - Alpha Decay Profile for UNCERTAINTY



Nervous (NERVOUS)

The NERVOUS signal captures elevated vocal tension, jitter, and stress markers—attributes that communication science links to cognitive load, discomfort, and defensive posture. As expected for a "high-bad" signal (direction = –1), its predictive pattern differs from CONFIDENCE and VDQ.

Performance Pattern

- D10 (least nervous) calls reliably outperform the benchmark, and
- D1 (most nervous) calls underperform, generating positive long–short (D10–D1) spreads.

Although the magnitude is slightly smaller than CONFIDENCE and VDQ, the pattern is highly consistent across horizons, especially H=5–15.

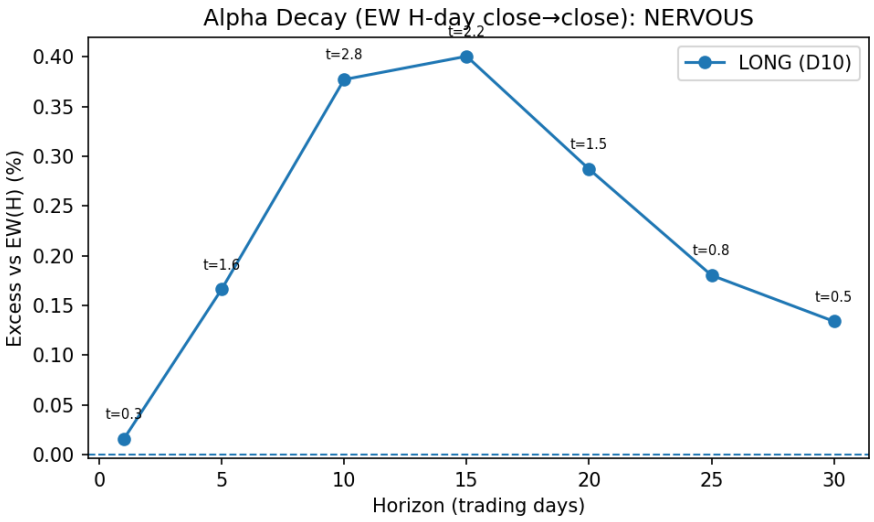
NERVOUS – Alpha Decay

The alpha-decay graph for NERVOUS shows:

- A strong early effect that peaks between H=5–10,
- Gradual decay toward H=30,
- Significance that is modest but persistent relative to other "high-bad" measures (e.g., AROUSAL).

The curve's shape is notable: it slopes gradually, not sharply, suggesting that investors slowly absorb vocal tension information, particularly when the Q&A covers more uncertain guidance topics.

Graph 8 – NERVOUS



5. Convergence and Divergence Analysis Between Text and Voice

Earnings calls convey information through two parallel channels: the content of language (what is said) and the paralinguistic vocal signal (how it is said). Prior literature in finance and communication science shows that these channels do not always move together, and when they diverge, the divergence itself can be informative. Executives may choose their words strategically, especially when discussing forward-looking conditions, but their vocal delivery is less consciously controlled and therefore more likely to reveal underlying conviction or uncertainty.

We therefore study whether markets respond differently when textual sentiment and vocal affect are aligned ("convergence") versus misaligned ("divergence").

- Convergence occurs when both text sentiment and vocal confidence point in the same direction. One may consider this combination 'double confirmation'.
    - *High–High Convergence* suggests strong conviction behind favorable messaging.
    - *Low–Low Convergence* suggests that negative or cautious messaging is sincere.
- Divergence occurs when the channels conflict.
    - *High Text Sentiment+ High Nervousness* reflects optimistic words delivered with vocal tension.
    - *Low Text Sentiment+ Low Nervousness* reflects negative content delivered calmly and without stress.

These patterns matter because investors often rely heavily on the transcript alone, and transcripts can obscure underlying sentiment, emotional state, or strategic tone-shaping. Voice offers an orthogonal signal that may reveal whether management is reassured, uncertain, or managing impressions.

Our objective is to test whether post-call excess returns differentiate these conditions in ways consistent with managerial conviction, information asymmetry, or selective communication.

5.1 Empirical Method

For each call, we compute:

- Textual sentiment (SENTIMENTPOLARITY)
- Vocal confidence and nervousness (CONFIDENCE, NERVOUS)

All features are ranked over the prior 90 days and we take the top/bottom quintile. We identify four conditions:

Condition	Text Sentiment	Voice Signal	Interpretation
Convergence (Hi–Hi, Positive)	High	High confidence	*Bullish message delivered confidently*
Convergence (Lo–Lo, Positive)	Low	Low confidence	*Bearish message delivered with concern*
Divergence (Hi–Hi, Negative)	High	High nervousness	*Upbeat words but stressed delivery*
Divergence (Lo–Lo, Negative)	Low	Low nervousness	*Negative messaging delivered calmly*

For Convergence and Divergence conditions, we measure the frequency of positive excess returns. We compare against the unconditional base rate at the same horizon and compute two-proportion z-statistics to assess significance.

5.2 Why This Matters

Executives have strong incentives and ample training to shape the content of their language. They have much less control over subtle prosodic elements such as pitch dispersion, jitter, articulation pressure, and micro-timing.

Thus:

- Text ≈ managerial intent (strategic)
- Voice ≈ managerial conviction (revealed)

Where these two signals converge, the message tends to be more credible, and markets react accordingly. Where they diverge, the gap itself can indicate hidden uncertainty, selective disclosure, or underlying confidence that is not fully reflected in the transcript.

This makes convergence and divergence a powerful framework for interpreting management tone as information, not noise.

Convergence (High Text Sentiment + High Vocal Confidence)

When bullish language is delivered with confident vocal tone, subsequent stock performance is marginally better than baseline. Most horizons show no statistically meaningful deviation from the base hit-rate, though a mild positive lift emerges at intermediate holding periods (10–15 days). Overall, markets appear to treat confident bullish messaging as expected rather than incrementally informative.

H	N	HitRate (%)	BaseRate (%)	Lift_vs_Base (pp)	z_vs_base
1	3614	49.5	49.4	0.14	0.164
5	3612	49.3	49.2	0.14	0.163
10	3614	50.9	49.5	1.39*	1.675
15	3614	51.5	49.7	1.79**	2.147
20	3614	49.9	49.7	0.15	0.184
30	3614	50.7	49.5	1.21	1.454

Convergence (Negative Sentiment Text + Low Vocal Confidence)

When executives deliver bearish wording with a vocal tone that also signals low confidence, forward returns weaken consistently. This paired signal of textual caution reinforced by audible concern tends to be treated by markets as credible and concerning. The result is a statistically significant degradation in hit-rates across most horizons, reflecting that aligned negative messaging is *significantly* more informative than hedged or mixed tones.

H	N	HitRate (%)	BaseRate (%)	Lift_vs_Base (pp)	z_vs_base
1	3050	49.0	49.4	-0.34	-0.38
5	3048	48.0	49.2	-1.23	-1.36
10	3048	47.8	49.5	-1.69*	-1.86
15	3048	47.5	49.7	-2.17**	-2.40
20	3048	47.7	49.7	-2.01**	-2.21
25	3048	47.4	49.5	-2.04**	-2.25
30	3048	46.0	49.0	-2.96**	-3.27

Divergence (Positive Text + High Nervousness)

When management expresses optimism in the script but sounds audibly nervous, markets do **not** treat the nervousness as a hidden negative signal. Across most horizons, hit rates exceed the base rate, with statistically strong lifts at 10–30 days. This pattern suggests investors discount the nervous tone and instead anchor to the optimistic guidance, leading to **better-than-baseline** forward performance despite the vocal stress.

H	N	HitRate (%)	BaseRate (%)	Lift_vs_Base (pp)	z_vs_base
5	3650	48.8	49.4	-0.57	-0.69
10	3650	52.1	49.2	2.86***	3.45
15	3650	51.5	49.5	1.93***	2.33

20	3650	52.2	49.7	2.48***	3.00
25	3650	52.0	49.7	2.29***	2.77
30	3650	51.3	49.5	1.86***	2.25

Divergence (Low Text Sentiment + Low Nervousness)

Calm delivery of negative information initially softens the blow, with short-horizon returns modestly outperforming the base rate. But as the holding period lengthens, the underlying negative news asserts itself: hit rates fall below baseline and become statistically significant at 10–15 days. Over longer windows the market recalibrates, interpreting this combination as **credible bad news**, with performance eventually drifting in line with or slightly below expectations.

H	N	HitRate (%)	BaseRate (%)	Lift_vs_Base (pp)	z_vs_base
5	3346	51.0	50.6	0.38	0.44
10	3346	48.7	50.8	-2.15***	-2.48
15	3346	48.9	50.5	-1.59*	-1.83
20	3346	50.6	50.3	0.28	0.32
25	3344	51.0	50.3	0.67	0.77
30	3344	52.0	50.5	1.52*	1.75

****99%, **95% *90* Confidence levels*

Collectively, the updated convergence and divergence results show that the interaction between *what* management says and *how* they say it conveys information that text alone does not capture. The market differentiates sharply between cases where verbal content and vocal affect are aligned versus in conflict, and prices the credibility of each communication channel differently.

In **convergence cases**, the market reacts asymmetrically. When bullish language is paired with confident vocal delivery, performance is largely indistinguishable from the base rate—suggesting that the market treats confident optimism as expected rather than incrementally informative. By contrast, when negative sentiment is delivered with low vocal confidence, forward returns deteriorate meaningfully and consistently, producing some of the strongest negative lifts in the sample. This pattern indicates that *aligned* caution—bearish text reinforced by an audibly concerned tone—is interpreted as highly credible and materially informative about future performance.

In **divergence cases**, the asymmetry reverses. Optimistic language delivered with audible nervousness does not produce the expected downside risk; instead, these firms outperform the baseline at most horizons. Investors appear to discount vocal stress when the verbal guidance remains strong, implying that nervousness alone is not treated as a reliable bearish signal. Conversely, when negative news is delivered calmly, short-horizon returns can be modestly positive, but this effect fades and then reverses. As the holding period extends, the underlying negative information dominates, and performance slips below the base rate. This suggests that vocal calm can delay—but not prevent—the market's eventual recognition of bad fundamentals.

Overall, the findings imply that **vocal cues act as a credibility filter**. When voice and text are aligned, the market places greater weight on the joint signal—rewarding calm positivity only marginally but penalizing sincere, audible caution sharply. When the channels diverge, investors

distinguish between strategic messaging and revealed conviction, generally siding with the textual message unless the divergence itself signals something about management's confidence. The transcript alone is incomplete; vocal delivery provides incremental information about belief strength and uncertainty that the market prices into returns over short and medium horizons.

Conclusion

This paper provides evidence that the paralinguistic channel of managerial communication or *how* executives speak during earnings calls, contains incremental, economically meaningful information that markets do not fully incorporate at the time of disclosure. Using a large-scale panel of CEO and CFO speech from Q&A sessions, combined with sentence-level acoustic extraction and an event-study framework aligned to strict, no-look-ahead trading rules, we show that vocal delivery systematically predicts short- and medium-horizon excess returns. Across seven distinct paralinguistic factors: including confidence, nervousness, delivery quality, PCA-based acoustic embeddings, and an integrated composite measure, the top-decile calls generate positive and statistically significant abnormal returns relative to a price-screened benchmark. These effects persist up to 30 trading days, consistent with post-call drift and market under-reaction to vocal cues.

By operating at the sentence level, we resolve a longstanding limitation in earnings-call voice research: call-level averages obscure meaningful variation in tone. Vocal affect only has economic interpretation relative to *what is being said* and *who is saying it*. Prepared remarks and Q&A differ fundamentally in informational structure; executives convey confidence, hesitation, and conviction unevenly across answers; CEO and CFO delivery patterns diverge systematically. Sentence-level segmentation allows us to compute hundreds of emotionally homogeneous micro-observations per call, enabling speaker standardization, capturing within-call variance, and substantially reducing noise through aggregation. These design choices make the resulting signals both more stable and more interpretable.

The empirical results show that markets respond asymmetrically to different types of vocal information. Positive vocal confidence and strong delivery quality are associated with upward drift following earnings calls, while elevated nervousness and high vocal entropy (UNCERTAINTY) predict weaker outcomes. Importantly, these effects are orthogonal to textual sentiment, underscoring the incompleteness of transcript-only analysis. The convergence–divergence framework further reveals that interactions between *what executives say* and *how they say it* carry distinct pricing implications. Bearish text delivered with vocal concern reliably predicts underperformance, while optimistic language delivered with vocal nervousness does not systematically signal downside risk, suggesting that investors partially discount nervous delivery when assessing upbeat messaging. Conversely, negative text delivered calmly (low nervousness) is eventually interpreted as high-conviction bad news, producing delayed underperformance.

Collectively, these findings demonstrate that the voice is not merely an ancillary channel but a material component of managerial disclosure. Paralinguistic cues encode conviction, stress, and internal belief states that are not always expressible, or intentionally expressed in text. The market partially processes this information on the event date but leaves a residual component that manifests as predictable post-call returns. For quantitative investors, this constitutes a distinct source of alpha, complementary to existing text-based approaches. For researchers, it suggests

that corporate communication should be studied as a multimodal system rather than a transcript-only artifact.

More broadly, the evidence points to a richer theory of disclosure in which managerial voice provides a real-time, behaviorally grounded signal of confidence and uncertainty—one that investors absorb gradually and sometimes imperfectly. As machine-listening technologies advance and as audio data becomes more systematically incorporated into the research pipeline, the informational role of paralinguistic cues is likely to grow, reshaping both academic models of communication and practical approaches to earnings-call analysis.

References

Alexopoulos, M., Han, X., Kryvtsov, O., & Zhang, X. (2024). More than words: Fed Chairs' communication during congressional testimonies. Journal of Monetary Economics, 142, 103515. https://doi.org/10.1016/j.jmoneco.2023.09.002

Amini, S., Hao, B., Yang, J., Karjadi, C., Kolachalama, V. B., Au, R., & Paschalidis, I. C. (2024). Prediction of Alzheimer's disease progression within 6 years using speech: A novel approach leveraging language models. Alzheimer's & Dementia. https://doi.org/10.1002/alz.13886

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, 70(3), 614–636. https://doi.org/10.1037/0022-3514.70.3.614

Baik, B., Kim, A. G., Kim, D. S., & Yoon, S. (2023). Managers' vocal delivery and real-time market reactions in earnings calls. SSRN Working Paper. https://ssrn.com/abstract=4398495

Biggiogera, J., Boateng, G., Hilpert, P., Vowels, M., Bodenmann, G., Neysari, M., Nussbeck, F., & Kowatsch, T. (2021). BERT meets LIWC: Exploring state-of-the-art language models for predicting communication behavior in couples' conflict interactions. arXiv:2106.01536.

Blankespoor, E., deHaan, E., & Zhu, C. (2020). How to talk when a machine is listening: Corporate disclosure in the age of AI. NBER Working Paper No. 27950. https://doi.org/10.3386/w27950

Brochet, F., Naranjo, P., & Yu, G. (2015). The capital market consequences of language barriers in the conference calls of non-U.S. firms. Journal of Financial Economics, 116(2), 404–426. (Working-paper version title often circulated as above.) https://ssrn.com/abstract=2154948

Call, A. C., Flam, R. W., Lee, J. A., & Sharp, N. Y. (2024). Managers' use of humor on public earnings conference calls. Review of Accounting Studies. (Advance/online first).

Chen, X. (L.), Levitan, S. I., Levine, M., Mandic, M., & Hirschberg, J. (2020). Acoustic-prosodic and lexical cues to deception and trust: Deciphering how people detect lies. Transactions of the Association for Computational Linguistics, 8, 199–214. https://doi.org/10.1162/tacl_a_00311

Fuchs, S., & Rochet-Capellan, A. (2021). The respiratory foundations of spoken language. Annual Review of Linguistics, 7, 13–30. https://doi.org/10.1146/annurev-linguistics-031720-103907

Gupta, P., et al. (2019). Bag-of-Lies: A multimodal dataset for deception detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

Hollien, H., Harnsberger, J. D., Martin, C. A., Hollien, K. A., & Alderman, T. M. (2014). Stress and deception in speech: Evaluation of layered voice analysis. Journal of Forensic Sciences, 59(2), 354–367. https://doi.org/10.1111/1556-4029.12338

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. https://doi.org/10.1111/j.1540-6261.2011.01705.x

Niederhoffer, K. G., & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. Journal of Language and Social Psychology, 21(4), 337–360. https://doi.org/10.1177/0261927X02021004003

Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., & Burzo, M. (2015). Deception detection using real-life trial data. In Proceedings of the 2015 ACM International Conference on Multimodal Interaction (ICMI '15) (pp. 59–66). https://doi.org/10.1145/2818346.2820744

Seibold, C., Wisotzky, E. L., Beckmann, A., Kossack, B., Hilsmann, A., & Eisert, P. (2025). High-quality deepfakes have a heart! Frontiers in Imaging, 4, 1504551. https://doi.org/10.3389/fimag.2025.1504551

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, 1994. https://doi.org/10.3389/fpsyg.2018.01994

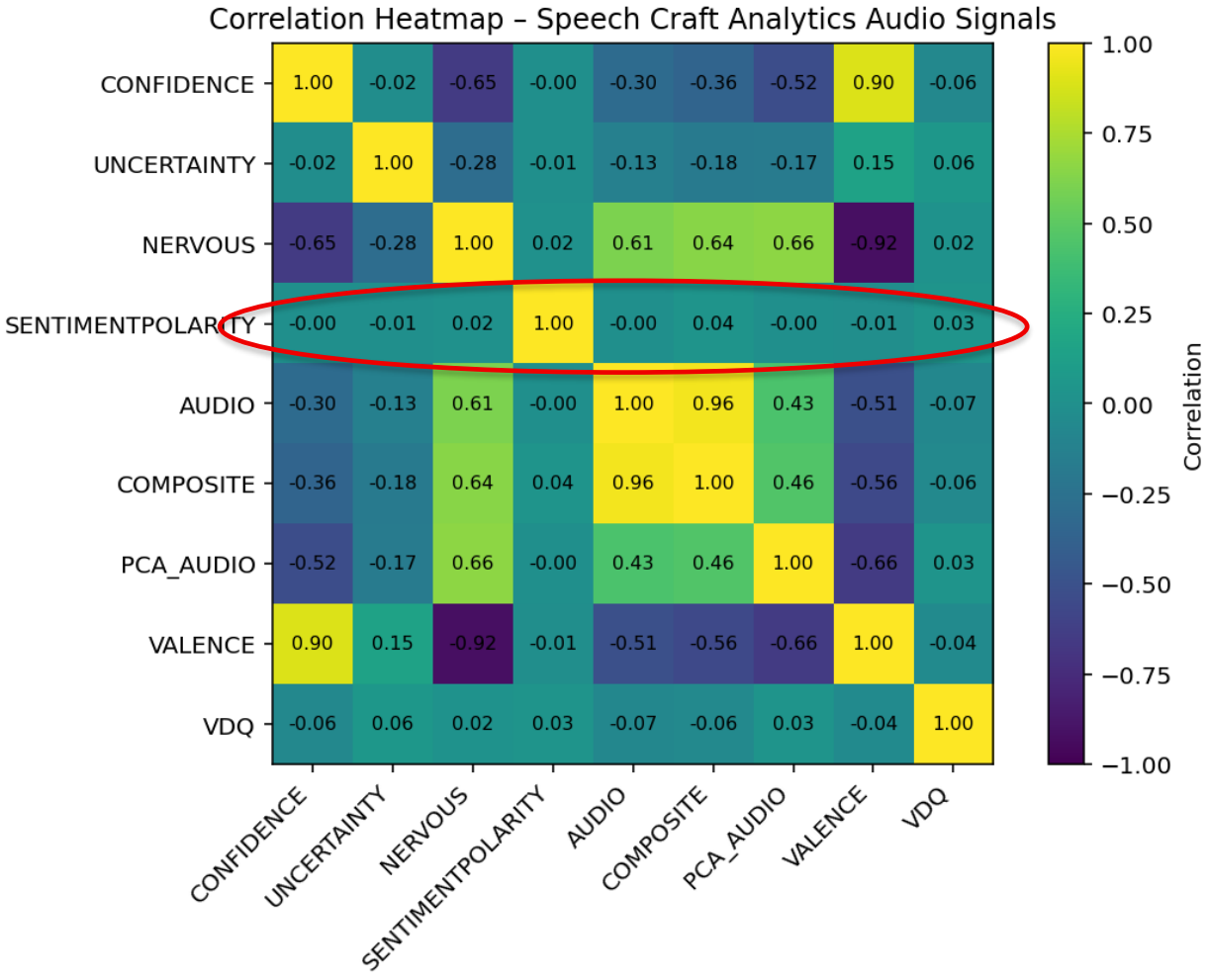
Wolfe Research (Luo's QES). (2023, June 14). The Leader's Voice: From transcripts to speech analysis in management presentations and conference calls. Wolfe Research LLC (industry report).

Rivolta, M., Minnick, K., et al. (2023). CEO–Outside Director ties and readability of financial reports. Working paper (ResearchGate distribution).

Speech Craft Analytics does not provide investment advice. Any information provided by Speech Craft Analytics, including data, analysis, reports, or any other communication, is solely for informational purposes and should not be considered a recommendation to invest, trade or engage in any financial transaction. Before making any financial decisions, users of Speech Craft Analytics services are advised to consult with a qualified professional for personalized advice. Speech Craft Analytics shall not be held liable for any actions or decisions taken by individuals or entities based on the information provided.

Appendix

1. Correlation Matrix of SCA Audio Signals + Sentiment Polarity



2. Yearly D10 Holding Period Average Excess Return