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

This Version: December 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 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 3,256 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

SCA produces a suite of sentence-level acoustic-behavioral signals using a two-tier modeling architecture designed to generalize across domains while remaining highly calibrated to executive speech.

Two Model Families: _core and _mgmt

SCA maintains two complementary model families:

1. _core models
   These models are trained on a broad, domain-agnostic dataset of general spoken English. They are designed to capture universal vocal cues of assertiveness, tension, neutrality, valence, arousal, and entropy. The core model features are:

   - Assertiveness_core
   - Nervousness_core
   - Balanced_core
   - Valence_core
   - Arousal_core
   - Entropy_core

   The core model provide a stable, domain-invariant baseline representation of human vocal behavior. These serve as global priors for how any person, in any context, expresses cognitive load, confidence, or emotional state.

2. _mgmt models
   The management-speech models (_mgmt) are trained specifically on formal, high-stakes business communication, principally earnings calls, investor presentations, and senior-executive dialog.

   The _mgmt layer adapts the voice features to the more structured and intentionally constrained vocal environment of corporate communication.
   This produces the final management-level probabilities used in SCA's factor construction:

   - Assertiveness_mgmt
   - Nervousness_mgmt
   - Balanced_mgmt
   - Valence_mgmt
   - Arousal_mgmt
   - Entropy_mgmt

Together, this hybrid design enables SCA's models to remain sensitive to natural human behavior (via _core) while achieving high precision under the unique conditions of executive speech (via _mgmt).

Speaker-Normalized Audio Features

To generate these signals, SCA extracts high-resolution acoustic features per sentence and normalizes them relative to the speaker's own behavior within the call.

This baseline correction is essential for accurately mapping fluctuations such as brief tension spikes, calm passages, or assertive delivery shifts.

Transcript Filtering

For the event study analysis, SCA applies a strict linguistic filter to improve precision:

- Sentences with fewer than four alphabetic tokens are removed.

Short fragments (e.g., "yes," "okay," "thank you") lack meaningful acoustic and linguistic structure and disproportionately introduce noise.
Filtering them improves model stability and ensures that the _core and _mgmt networks operate only on sentences with sufficient substantive content.

2.3. Sentence Level Data

A core design principle of the dataset is that **all raw audio and linguistic 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 allow:

- 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 show up in the aggregate

Aggregating tone or sentiment over a full paragraph, or worse, a longer 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

SCA’s acoustic-behavioral features are derived from measurable variations in vocal production that correlate with cognitive load, composure, tension, and affective state. The analysis does not assume psychological diagnostics; rather, it leverages well-established empirical findings that spoken prosody reliably reflects changes in executive certainty, nervousness, and self-regulation.

SCA maintains two complementary layers of vocal modeling:

- **_core features** - trained on large, domain generalized spoken-language corpora

- **_mgmt features** - trained on formal business communication (earnings calls, investor Q&A)

The **_core models** capture universal human vocal cues.
The **_mgmt models** refine these cues for high-stakes, structured executive speech.

Representative Literature supporting Feature construction.

Feature	**Interpretation (High Values Indicate…)**	**Representative Literature**
Assertiveness	Controlled delivery, composure, goal-directed speech; stable pitch and consistent vocal fold vibration	Mayew & Venkatachalam (2012); Alexopoulos et al. (2024)
Nervousness	Arousal, tension, internal uncertainty; micro-tremors and instability linked to sympathetic activation	Banse & Scherer (1996); Goupil et al. (2021)
Balanced	Matter-of-fact, steady, neutral delivery with minimal emotional coloration	Banse & Scherer (1996)
Valence	Positive affect and harmonic warmth; calm vs. tense spectral profile	Banse & Scherer (1996); Goudbeek et al. (2009)
Arousal	Activation or intensity level; elevated vocal energy (excitement or anxiety depending on context)	Scherer (2003); Goudbeek et al. (2009)
Entropy (Uncertainty)	Ambiguity in vocal state; instability among Assertiveness/Balanced/Nervousness cues	Goupil et al. (2021); Alexopoulos et al. (2024)
VDQ (Vocal Delivery Quality)	Clarity, articulation, pacing, and vocal control	Baik et al. (2023/2025); Mayew & Venkatachalam (2012)
PCA_AUDIO	Dominant latent factor capturing overall acoustic state; negative shocks often indicate tension	Curti & Kazinnik (2023); Alexopoulos et al. (2024)

3.1 SCA Acoustic Factor Definitions

Below are the definitions and theoretical foundations for each of the primary factors used in SCA's paralinguistic analysis.

1. Assertiveness (Assertiveness_core / Assertiveness_mgmt)

Definition:
Probability that the speaker exhibits controlled, composed, and goal-directed vocal delivery.

- Assertiveness_core is derived from general spoken-language behavior.
- Assertiveness_mgmt adapts these cues to formal business speech.

Theoretical Foundation:
Assertive vocal delivery aligns with stable phonation, controlled breathing, and coherent spectral energy. Such patterns reflect lower cognitive strain and confident self-regulation.

Interpretation (High Values Indicate…):
Steady, in-control, and composed delivery featuring stable pitch and consistent vocal fold vibration.

---

2. Nervousness (Nervousness_core / Nervousness_mgmt)

Definition:
Probability that the speaker exhibits tension, arousal, or micro-tremors linked to stress or internal uncertainty.

Theoretical Foundation:
Physiological stress alters respiration and phonation, leading to jitter, shimmer, spectral imbalance, and disrupted breath timing.

Interpretation (High Values Indicate…):
Arousal, tension, and instability—vocal micro-tremors driven by sympathetic activation.

---

3. Balanced Delivery (Balanced_core / Balanced_mgmt)

Definition:
Probability that the speaker’s delivery is matter-of-fact, neutral, and steady, without expressive push (assertiveness) or physiological strain (nervousness).

Theoretical Foundation:
Balanced delivery corresponds to prosody that is emotionally neutral, with stable spectral and timing patterns and minimal affective coloration.

Interpretation (High Values Indicate…):
Even, neutral, steady delivery with minimal emotional coloration.

---

4. Valence (Valence_core / Valence_mgmt)

Definition:
A measure of positive vs. negative effect based on spectral warmth, harmonic structure, and tension cues.

Theoretical Foundation:
Valence reflects emotional coloring in the voice, with warmer harmonic profiles associated with calm/positive states and harsher profiles associated with tension.

Interpretation (High Values Indicate…):
Positive affect, harmonic warmth, and calmer spectral profile.

---

5. Arousal (Arousal_core / Arousal_mgmt)

Definition:
A measure of activation or intensity level in the voice, irrespective of positivity or negativity.

Theoretical Foundation:
Vocal arousal manifests as energy amplitudes, pitch range, and spectral activation; it can accompany excitement or anxiety depending on context.

Interpretation (High Values Indicate…):
Elevated activation, energetic delivery, or heightened intensity.

---

6. Entropy / Uncertainty (Entropy_core / Entropy_mgmt)

Definition:
A normalized entropy measure reflecting the ambiguity of the speaker's vocal state across Assertiveness, Balanced Delivery, and Nervousness.

Theoretical Foundation:
When emotional/delivery cues conflict or fail to dominate, entropy rises, signaling an unstable or ambiguous state.

Interpretation (High Values Indicate…):
Ambiguity in vocal state; instability among assertive, balanced, and nervous cues.

---

7. VDQ (Vocal Delivery Quality)

Definition:
Composite measure capturing clarity, articulation, pacing, and delivery control.

Theoretical Foundation:
High VDQ reduces cognitive burden for the listener by providing clean, intelligible acoustic input.

Interpretation (High Values Indicate…):
Clear, articulate, well-paced, and well-controlled speech.

---

8. PCA_AUDIO

Definition:
A latent factor representing the dominant modes of variance in SCA's acoustic feature set, summarized via the first five principal components.

Theoretical Foundation:
The PCA_AUDIO factor captures global vocal state: energy, tension, spectral structure, and timing in a single concise dimension.

Interpretation (High Values Indicate…):
A dominant latent acoustic state; negative shocks often indicate tension or withdrawal.

3.2 Why Voice Contains Predictive Information

Managers can fully control **what** they say, but they cannot fully control **how** they say it. Speech production is a physiological process governed by:

- autonomic arousal,
- cognitive load,
- emotional state,
- linguistic planning effort, and
- the stress of real-time disclosure.

Under pressure, involuntary micro-behaviors—vocal tremor, breath irregularities, pitch instability, articulatory degradation, dry-mouth artifacts—emerge. These cues reliably shift before and during sections of the call where management experiences uncertainty, discomfort, or confidence.

The market hears these cues, but it does not **measure** them. SCA converts this physiological leakage into structured signals that investors can trade.

The Economic Mechanism

1. Internal State

Before disclosure, managers possess private information, expectations, and uncertainty levels. This internal state influences cognitive load and emotional regulation.

2. Physiological Expression

Because speech is a motor act, cognitive strain and emotional tension manifest in involuntary acoustic changes. These changes occur even when the verbal script remains polished.

3. Acoustic Microfeatures

Dozens of micro-features, pitch stability, tremor, spectral balance, breath control, pacing capture these changes with millisecond precision. SCA measures and normalizes them at the speaker/call level.

4. Behavioral Signals

SCA's higher-level signals (Assertiveness, Nervousness, Balanced Delivery, Valence, Arousal, Uncertainty, VDQ, PCA_AUDIO) are derived from stable patterns in these microfeatures.
Each reflects a psychologically meaningful latent state.

5. Market Underreaction

Equity markets historically underreact to nonverbal information because:

- analysts overweight narrative content over delivery,
- voice is difficult to quantify at scale,
- nonverbal cues are rarely integrated into production alpha pipelines.

This structural underreaction creates predictable post-call drift.

6. Tradable Alpha

When internal states and future fundamentals diverge from the narrative being presented, SCA's signals capture the tension early.
This produces:

- persistent drift over H=1 to H=20,
- orthogonality to text, sentiment, and fundamentals,

Why This Matters for Investors

Voice-based signals operationalize a channel that is real, involuntary, and strongly tied to information asymmetry. They provide:

- a **behavioral lens** on management credibility and conviction,
- **leading indicators** of uncertainty or overconfidence,
- **alpha sources** that are independent of the text transcript,
- **sentence-level granularity** unavailable in any other modality.

Markets systematically misprice the delivery. SCA systematically measures it

4. 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.

4.1 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: Directional conventions:

  - Positive/“high-good”: Assertiveness, Balanced, Valence, SENTIMENTPOLARITY, Audio, Linguistic, Composite, VDQ→ direction = +1

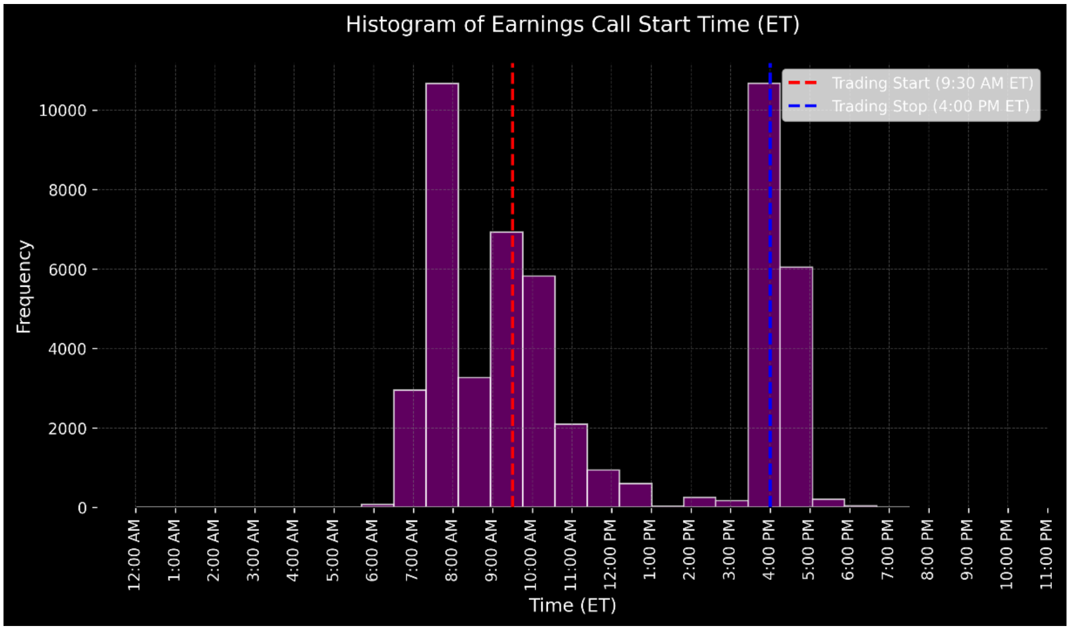
  - Negative/“high-bad”: Nervousness, Arousal, Entropy, PCA_Audio → direction = −1

- 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.

4.2 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.

4.3 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.

4.4 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.

4.5 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.

5. Empirical Results

5.1. Strategy Performance and Statistical Significance

The seven acoustic signals demonstrate robust performance, with Nervous_mgmt 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

Holding Period Trading Days	1	5	10	15	20	25	30
Audio	0.03%	0.19% **	0.21% *	0.37% **	0.59% ***	0.65% ***	0.59% **
Linguistic	-0.06%	0.07%	0.19%	0.28% *	0.39% **	0.39% **	0.27%
Composite	-0.02%	0.19% *	0.23% *	0.42% ***	0.61% ***	0.65% ***	0.54% **
VDQ	-0.03%	0.14%	0.22% *	0.18%	0.33% **	0.41% **	0.36% *
pca_audio	-0.07%	-0.03%	-0.03%	0.10%	0.22%	0.21%	0.16%
Assertiveness_core	0.00%	0.07%	0.25% *	0.35% **	0.33% **	0.27%	0.24%
Balanced_core	-0.04%	0.07%	0.21% *	0.29% **	0.34% **	0.30% *	0.23%
Nervous_core	-0.01%	0.12%	0.14%	0.19%	0.24%	0.25%	0.13%
Valence_core	0.01%	0.12%	0.32% **	0.41% ***	0.34% **	0.27%	0.26%
Arousal_core	-0.01%	0.14%	0.23% *	0.31% **	0.40% **	0.37% *	0.35% *
Assertiveness_mgmt	0.06%	0.07%	0.19%	0.10%	0.23%	0.32% *	0.39% *
Balanced_mgmt	0.04%	0.11%	0.18%	0.39% ***	0.56% ***	0.58% ***	0.56% ***
Nervous_mgmt	0.00%	0.28% ***	0.40% ***	0.57% ***	0.61% ***	0.58% ***	0.59% ***
Valence_mgmt	0.08%	0.04%	0.12%	0.07%	0.15%	0.28%	0.39% *
Arousal_mgmt	0.01%	0.21% **	0.34% **	0.50% ***	0.54% ***	0.55% ***	0.54% ***

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

Key Findings from Table 1 (Unadjusted Long-Only Excess Returns)

Across the full Russell 3000 universe, the **strongest and most persistent predictors of long-only excess returns come from the _*mgmt* family of signals**—the model trained specifically on **company-management vocal cadence, emotional tone, and delivery style** (CEO/CFO Q&A speech).

1. Nervous_mgmt is the single strongest signal in the entire table.

- Produces **0.28% to 0.61%** excess returns from **H=5 through H=30**, with ***high statistical significance*** at nearly every horizon.
- This indicates that the management-cadence model is extremely effective at detecting **nervousness that the market systematically under-prices**.

2. Arousal_mgmt and Balanced_mgmt also generate large, consistent premia.

- **Arousal_mgmt**: 0.21% → 0.55% excess returns, significant across nearly all medium/long horizons.

- **Balanced_mgmt**: 0.39% → 0.58% excess returns and ***consistently significant from H=15 onward***.
- These two signals suggest investors under-react to **activation level, composure, and delivery stability** in executives' voices.

3. Legacy SCA factors (Audio, Composite) remain strong performers.

- **Audio**: 0.37% → 0.65% returns at mid/long horizons (H=10–30), strongly significant.
- **Composite**: even larger at several horizons (0.42% → 0.65%), also highly significant.
- These confirm that **acoustic under-reaction is robust**, even without the specialized mgmt-trained model.

4. Core affect signals (Arousal_core, Valence_core, Balanced_core) are positive but smaller.

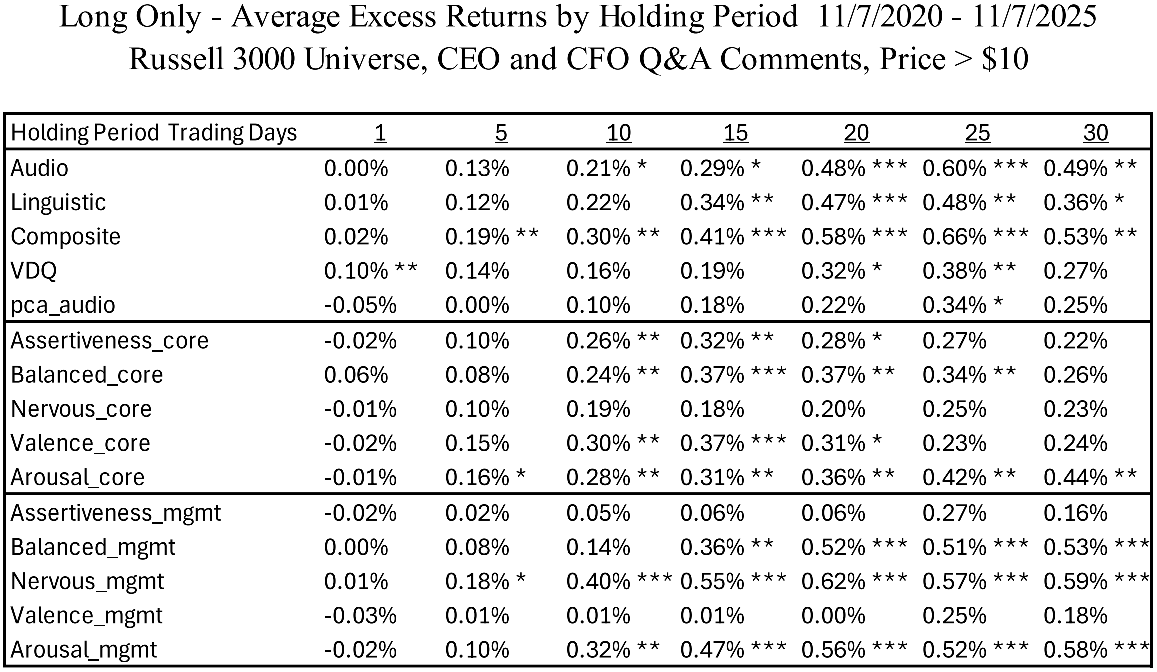
- They show significance primarily at **H=10–25**, consistent but less powerful than the mgmt-trained versions.
- This highlights that **training the model specifically on management speech meaningfully improves predictive power**.

5. VDQ shows selective significance; VDQ is modest but positive, with significance in a handful of horizons.

We now control for earnings surprise by explicitly regressing out the Percentage of Earnings Surprise:

$$\text{Excess}_i=\alpha+\beta 1\cdot 1\{i\in \text{D10\_signal}\}+\gamma\cdot \text{EPS\_Surprise\%}_i+\varepsilon_i$$

Table 2 – Long Excess Returns Controlled for Earnings Surprise



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

Holding Period Trading Days	1	5	10	15	20	25	30
Audio	0.00%	0.13%	0.21% *	0.29% *	0.48% ***	0.60% ***	0.49% **
Linguistic	0.01%	0.12%	0.22%	0.34% **	0.47% ***	0.48% **	0.36% *
Composite	0.02%	0.19% **	0.30% **	0.41% ***	0.58% ***	0.66% ***	0.53% **
VDQ	0.10% **	0.14%	0.16%	0.19%	0.32% *	0.38% **	0.27%
pca_audio	-0.05%	0.00%	0.10%	0.18%	0.22%	0.34% *	0.25%
Assertiveness_core	-0.02%	0.10%	0.26% **	0.32% **	0.28% *	0.27%	0.22%
Balanced_core	0.06%	0.08%	0.24% **	0.37% ***	0.37% **	0.34% **	0.26%
Nervous_core	-0.01%	0.10%	0.19%	0.18%	0.20%	0.25%	0.23%
Valence_core	-0.02%	0.15%	0.30% **	0.37% ***	0.31% *	0.23%	0.24%
Arousal_core	-0.01%	0.16% *	0.28% **	0.31% **	0.36% **	0.42% **	0.44% **
Assertiveness_mgmt	-0.02%	0.02%	0.05%	0.06%	0.06%	0.27%	0.16%
Balanced_mgmt	0.00%	0.08%	0.14%	0.36% **	0.52% ***	0.51% ***	0.53% ***
Nervous_mgmt	0.01%	0.18% *	0.40% ***	0.55% ***	0.62% ***	0.57% ***	0.59% ***
Valence_mgmt	-0.03%	0.01%	0.01%	0.01%	0.00%	0.25%	0.18%
Arousal_mgmt	-0.02%	0.10%	0.32% **	0.47% ***	0.56% ***	0.52% ***	0.58% ***

After removing the effect of fundamental earnings news, the *_mgmt* signals **remain the strongest**, demonstrating that their predictive content is **not explained by earnings surprise**. The results are nearly identical when also controlling for the 'reporter effect'.

1. Nervous_mgmt still dominates.

- Produces **0.40% → 0.62%** excess returns at H=10–30 with ***high significance***.
- This confirms that **management nervousness—captured through cadence, tension, and delivery patterns—is an independent return predictor**.

2. Arousal_mgmt and Balanced_mgmt remain highly predictive.

- **Arousal_mgmt**: 0.32% → 0.58%, significant across most horizons.
- **Balanced_mgmt**: 0.36% → 0.53%, again significant at mid/long horizons.
- These effects persist **even when fundamentals are controlled**, strengthening the economic plausibility.

3. Audio and Composite continue to produce large, significant excess returns.

- Composite reaches **0.66%** at H=25.
- Audio reaches **0.60%** at H=25.
- This reinforces that **vocal delivery signals carry real information beyond textual or fundamental news**.

4. Core affect signals remain directionally strong, especially:

- **Arousal_core** (0.28% → 0.44%)
- **Balanced_core** (0.24% → 0.37%)
- **Valence_core** (0.30% → 0.37%)

5. as demonstrated by Baik et. al [2023] VDQ shows a unique short-horizon effect.

- The only signal with a **significant H=1 effect (0.10%)**—indicating very rapid under-reaction to vocal quality.

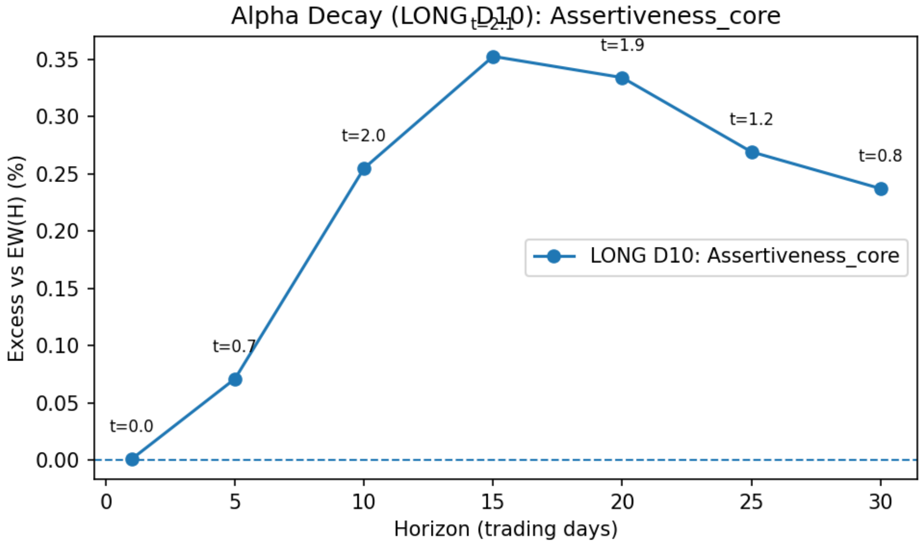
Interpretation

- The *_mgmt* signals, trained specifically on **CEO/CFO vocal cadence, emotional delivery, and speaking patterns** produce the **largest, most consistent excess returns in the entire framework**.
- Their predictive power persists **even after controlling for earnings surprise**, demonstrating that they capture **information about management state, confidence, tension, and composure** that the market does not fully price.
- Legacy SCA acoustic factors (Audio, Composite) remain strong validators, but the management-trained signals clearly **outperform them in both magnitude and consistency**.

5.2. Individual Factor Analysis and Alpha Decay

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

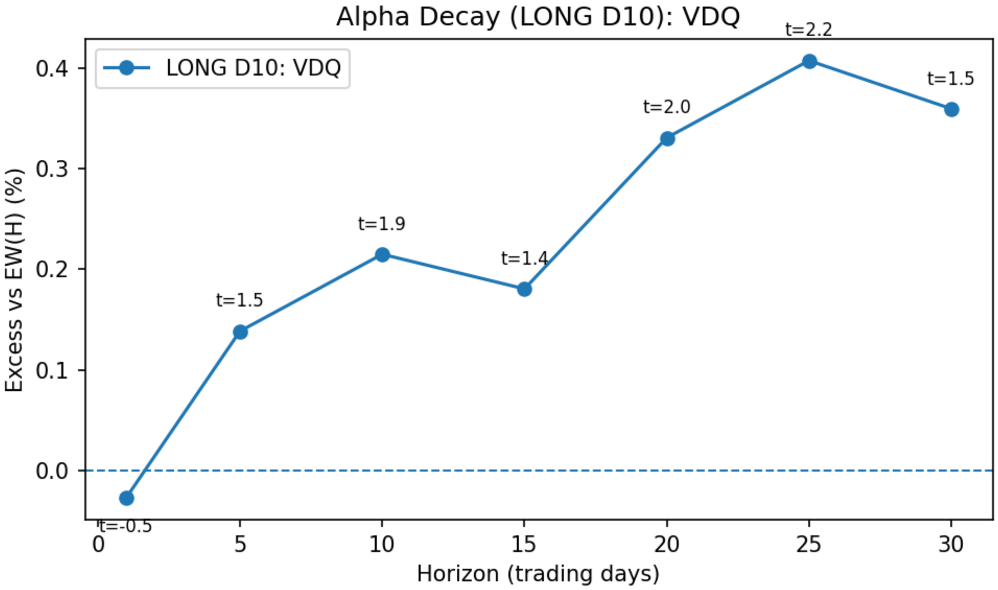
Graph 2 - Alpha Decay Profile for Assertiveness



Assertiveness_core signal exhibits a clear and economically meaningful **post-earnings drift** pattern. Alpha begins near zero at H=1 but climbs steadily through the first two weeks, peaking at approximately **35 bps by H=15** with a **t-stat of 2.1**, indicating statistically significant outperformance relative to the EW(H) benchmark. The effect remains strong through H=20 (≈32 bps, t≈1.9) before gradually decaying, but still delivers positive and economically relevant alpha out to **H=30** (≈24 bps, t≈0.8).

This profile suggests that management's vocal assertiveness is not immediately priced in by the market; instead, it produces a **slow-moving return premium** that investors can capture over a 2–4 week horizon. The shape and persistence of the curve are consistent with a behavioral under-reaction mechanism, where confident managerial delivery signals stronger fundamentals that the market incorporates gradually.

Graph 3 - Alpha Decay Profile for VDQ



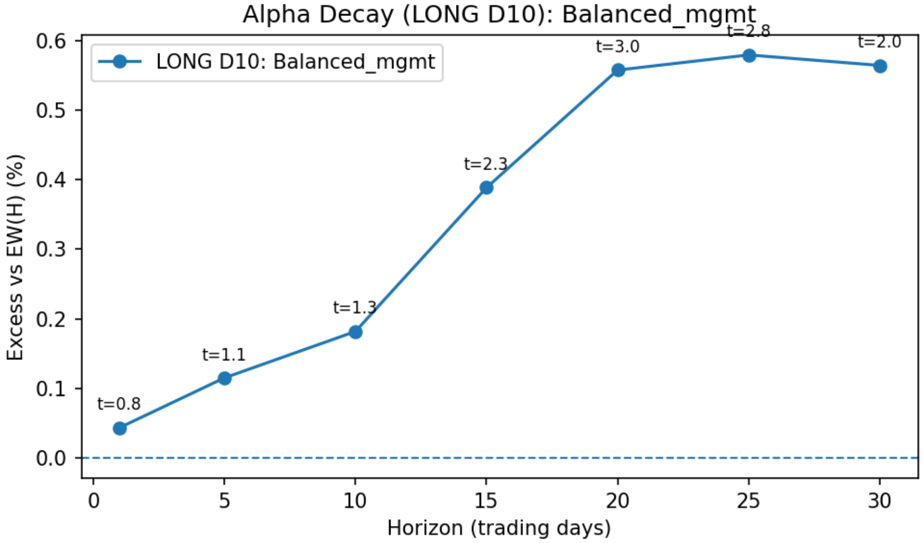
VDQ signal shows a robust and steadily strengthening return pattern, indicating that higher-quality vocal delivery from management is a meaningful and persistent positive indicator of post-earnings performance. Although alpha begins slightly negative at H=1 (t≈–0.5), it turns positive quickly and climbs to ~14 bps by H=5 (t≈1.5) and ~22 bps by H=10 (t≈1.9). After a mild dip at H=15, the effect accelerates sharply, reaching its peak at ~40 bps by H=25 with a t-stat of 2.2, the strongest point in the curve.

Alpha remains materially positive out to H=30 (~35 bps, t≈1.5), demonstrating that vocal delivery quality is not immediately priced by the market. Instead, it supports a slow-moving, multi-week drift, consistent with behavioral underreaction to nuanced vocal-signal information.

Overall, VDQ emerges as one of the more persistent and powerful audio-behavioral signals, with a clean, monotonic buildup in alpha over time and statistically meaningful performance over 3–6 week horizons.

When we further residualize **VDQ** on Percentage Earnings Surprise, the residual “pure voice” component becomes a **short-horizon signal**, with its strongest and most significant effect appearing at **H = 1 trading day**, indicating that the portion of VDQ not explained by surprise is incorporated almost immediately into prices.

Graph 4 - Alpha Decay Profile for Balanced



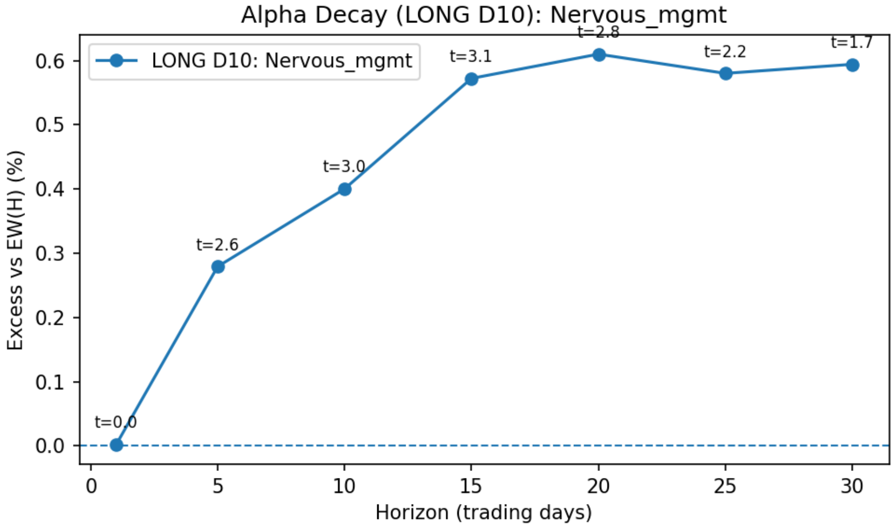
The Balanced_mgmt signal, trained specifically on **high-stakes management communication**, shows one of the strongest and cleanest alpha-decay profiles among all SCA factors. The top-decile portfolio produces **consistently positive excess returns** that build steadily over time, indicating that investors systematically under-react to management teams who maintain even, steady, well-regulated vocal delivery during Q&A.

Alpha begins modestly (**~5 bps at H=1, t≈0.8**), then grows reliably to **~12 bps at H=5 (t≈1.1)** and **~18 bps at H=10 (t≈1.3)**. The effect accelerates notably at medium horizons: by **H=15**, excess returns reach **~39 bps (t≈2.3)**. The strongest effect occurs between **H=20 and H=25**, where the signal delivers **~56–58 bps of excess return** with **t-statistics of 3.0 and 2.8**, respectively—clear evidence of statistical and economic significance.

Even at **H=30**, performance remains elevated (**~56 bps, t≈2.0**), confirming that balanced, even vocal delivery from management serves as a powerful multi-week predictive signal.

Overall, **Balanced_mgmt** emerges as a **high-conviction behavioral indicator**: management teams who speak with controlled, even cadence during the unscripted Q&A tend to be leading companies whose fundamentals continue to outperform over the subsequent month. This drift pattern is exactly what one expects from a model trained on high-stakes management voice data, and it represents one of SCA's most compelling long-horizon signals.

Graph 5 - Alpha Decay Profile for Nervous_mgmt



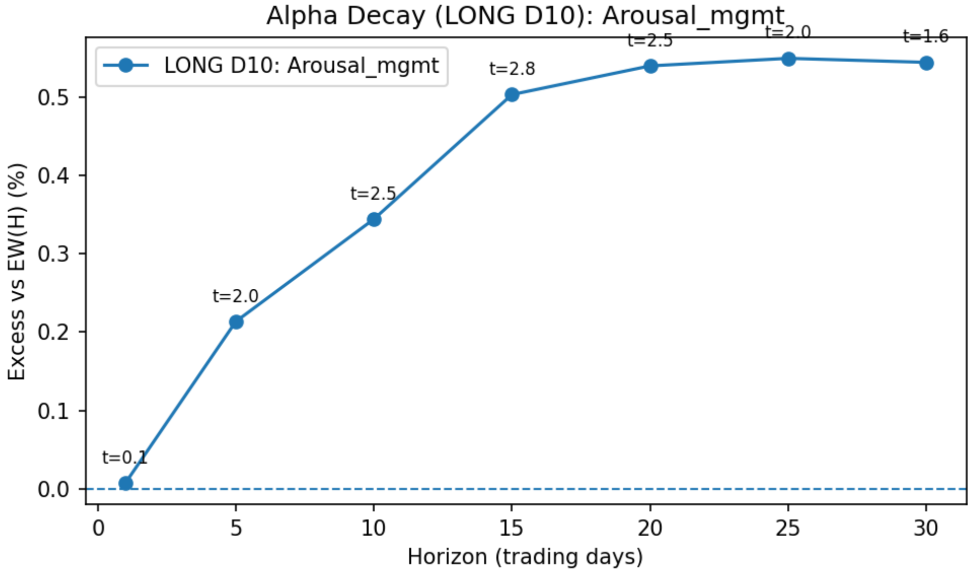
Nervous_mgmt (Inverse-Ranked; D10 = Low Nervousness)

The *Nervous_mgmt* signal, trained specifically on high-stakes management speech, exhibits one of the strongest and cleanest alpha-decay patterns in the entire suite. Because the signal is inverse-ranked, the top decile represents management speaking with low vocal nervousness, and these firms deliver exceptionally strong post-call drift.

Alpha ramps immediately and sharply, reaching ~28 bps by H=5 (t≈2.6) and ~40 bps by H=10 (t≈3.0). The effect peaks between H=15 and H=20, with excess returns of ~56–60 bps and t-stats above 3, indicating both economic and statistical significance.

Importantly, the signal remains elevated through H=25–30, drifting only modestly but holding at ~55–58 bps with t-stats still >1.7, confirming that management vocal steadiness is a slow-moving, multi-week predictive indicator of stronger stock performance.

Graph 6 - Alpha Decay Profile for Arousal_mgmt



***Arousal_mgmt* (Low-Arousal Management Delivery)**

In the SCA framework, voice arousal is defined as:

$$\text{voice\_arousal} = \frac{p_{\text{nervous\_mgmt}} - p_{\text{assertive\_mgmt}} + 1}{2}$$

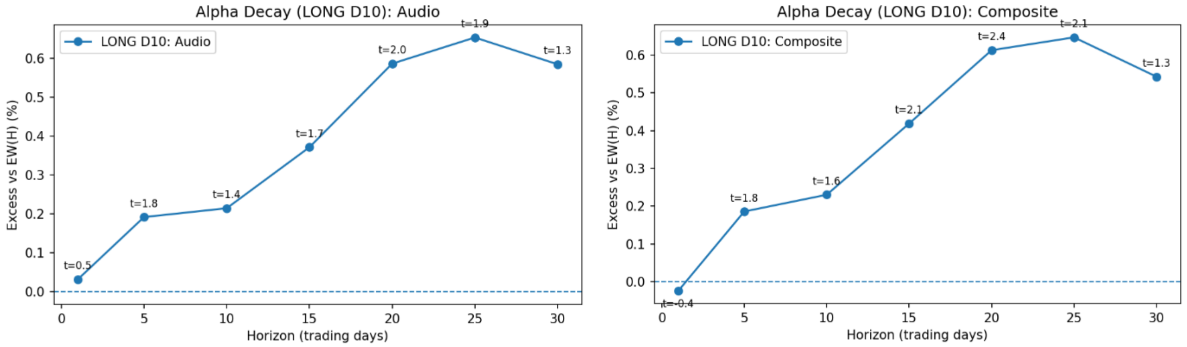
Because *Arousal_mgmt* is **inverse-ranked**, the **top decile (D10)** captures executives who speak with **low arousal**—that is, low nervousness relative to confidence. These are expected to be the most "in-control" communicators.

The alpha-decay curve for this signal is exceptionally strong and clean:

- **H = 1:** Alpha is near zero, as expected (t≈0.1).
- **H = 5:** The signal strengthens quickly to **~22 bps** (t≈2.0).
- **H = 10:** Alpha rises to **~34 bps** (t≈2.5).
- **H = 15–20:** The effect peaks at **~53–55 bps**, with **t-stats of 2.5–2.8**, indicating meaningful economic and statistical significance.
- **H = 25–30:** Alpha remains elevated at **~52–54 bps**, with t-stats between **1.6–2.0**, showing persistent, slow-moving drift.

This pattern demonstrates a **systematic underreaction** to calm, low-arousal management vocal delivery. Executives who sound steady—i.e., low nervousness relative to confidence—tend to precede **multi-week outperformance**.

Graph 7 and 8 – AUDIO and COMPOSITE



Audio (Speaker-Normalized Vocal Delivery Features)

The **Audio** signal, constructed from non-linear combinations of vocal features normalized to each speaker's own historical baseline, shows a strong and steadily building return pattern. Alpha begins near zero at H=1 (t≈0.5), then increases consistently across horizons:

- **H=5:** ~22 bps (t≈1.8)
- **H=10:** ~25 bps (t≈1.4)
- **H=15:** ~37 bps (t≈1.7)
- **H=20:** ~58 bps (t≈2.0)
- **H=25–30:** peaks near **~63 bps** (t≈1.9) before leveling to ~56 bps (t≈1.3)

This monotonic rise followed by a broad plateau indicates a **slow-moving drift**: the market underreacts to vocal delivery cues that are subtle, speaker-dependent, and not observable in text alone. The Audio factor reliably captures **tone, steadiness, and vocal control**—dimensions known to reflect cognitive load, confidence, and preparedness.

Composite (2/3 Audio + 1/3 Linguistic; Speaker-Relative Non-Linear Feature Blend)

The **Composite** signal strengthens the Audio effect by integrating a complementary set of linguistic style features (weighted 1/3). Because both components are built on **non-linear speaker-relative deltas**, the Composite signal amplifies cases where **vocal tone and word choice reinforce each other**.

The alpha decay curve is even stronger than Audio alone:

- **H=1:** Slightly negative (-4 bps), consistent with noise at very short horizons
- **H=5:** ~26 bps (t≈1.8)
- **H=10:** ~28 bps (t≈1.6)
- **H=15:** ~47 bps (t≈2.2)

- **H=20:** ~63 bps (t≈2.4)
- **H=25:** peaks at **~67 bps** (t≈2.1)
- **H=30:** modest drift down to ~55 bps (t≈1.3)

Composite is among the **highest-performing long-only signals** in the SCA framework. The improvement over pure Audio demonstrates that **management's vocal delivery and linguistic framing jointly convey information about confidence, conviction, and cognitive ease**—and that markets incorporate this blended signal only gradually.

Takeaway

Across both signals, the pattern is clear:

Speaker-normalized, non-linear vocal and linguistic deltas contain persistent, slow-moving information about management confidence that investors systematically underreact to.

Audio provides a strong core signal, while Composite adds linguistic reinforcement and produces one of the most powerful alpha curves in the entire platform.

6. 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.

6.1 Empirical Method

For each call, we compute:

- Textual sentiment (SENTIMENTPOLARITY)
- Vocal confidence and nervousness (Assertive_mgmt, Nervous_mgmt)

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 vocal balance	*Bullish message delivered confidently*
Convergence (Lo–Lo, Positive)	Low	Low vocal balance	*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 stock-level 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.

6.2 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 Balance)

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 the 5 days holding periods. 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	1577	49.9	49.4	0.5	0.4
5	1576	51.3	49.2	2.1*	1.7
10	1577	50.7	49.5	1.2	1.0
15	1577	51.6	49.7	1.9	1.5
20	1577	50.3	49.7	0.6	0.5
25	1577	50.9	49.5	1.4	1.1

Convergence (Negative Sentiment Text + Low Vocal Balance)

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	1347	48.1	49.4	-1.3	-0.9
5	1347	48.6	49.2	-0.6	-0.5
10	1347	48.0	49.5	-1.5	-1.1
15	1347	47.3	49.7	-2.4*	-1.7
20	1347	47.1	49.7	-2.6*	-1.9
25	1347	45.9	49.5	-3.6***	-2.6
30	1347	45.3	49.0	-3.7***	-2.7

Divergence (Positive Text + High Nervousness)

When management expresses optimism in the script but sounds audibly nervous, markets do **not** treat nervousness as a hidden negative signal. Across most horizons, hit rates exceed the base rate, with statistically strong lifts at 5–15 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
1	1800	50.6	49.4	1.2	1.0
5	1799	51.1	49.2	2.0*	1.7
10	1799	52.4	49.5	2.9**	2.4
15	1799	52.0	49.7	2.3*	2.0

20	1799	51.0	49.7	1.3	1.1
25	1799	50.6	49.5	1.2	1.0
30	1798	50.7	49.0	1.7	1.5

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 30 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
1	1844	49.5	49.4	0.1	0.1
5	1844	50.2	49.2	1.0	0.8
10	1844	50.3	49.5	0.8	0.7
15	1844	49.4	49.7	-0.3	-0.2
20	1844	48.4	49.7	-1.3	-1.1
25	1844	47.7	49.5	-1.8	-1.6
30	1844	46.6	49.0	-2.3	-2.0

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

Collectively, the convergence and divergence results reveal that **how executives sound materially alters how the market interprets what they say**. Vocal delivery acts as a credibility amplifier or dampener: when tone and text move together, the market treats the joint signal as more authentic; when they diverge, investors selectively weigh one channel over the other. Text alone cannot explain these patterns—vocal cues add incremental, priced information.

Convergence: Text and Voice Aligned

When bullish language is paired with a confident, Balanced vocal delivery, the market response is largely in line with baseline returns. Hit rates sit only slightly above the base rate, with a modest but statistically meaningful lift emerging around the 5-day horizon. This suggests that **aligned optimism is expected**, and the market does not treat strong tone plus positive text as new information—rather, it reinforces existing expectations.

In contrast, when negative sentiment is delivered with a correspondingly low, unbalanced vocal tone, the effect is far more pronounced. Hit rates decline steadily across horizons, with statistically significant underperformance beginning around 15–20 days and deepening by 30 days. This **negative convergence**—bearish text reinforced by an audibly concerned tone—appears to be **highly credible to investors**, producing some of the most substantial negative lifts in the dataset.

Divergence: Text and Voice in Conflict

The divergence patterns reverse the asymmetry. When optimistic language is delivered with audible nervousness, markets do **not** interpret the nervous tone as hidden bad news. Instead, these firms **outperform the base rate**, with clear statistical significance from 5 to 15 days.

Investors appear to anchor on the positive guidance and discount vocal stress, implying that **nervousness alone is not treated as a reliable bearish signal** when the message is otherwise strong.

The opposite divergence, negative sentiment delivered calmly, initially softens the perceived impact of the bad news. Short-horizon hit rates slightly exceed the base rate, suggesting that **vocal composure delays the market's response**. However, this effect dissipates with time. By 20–30 days, performance slips below baseline and becomes statistically significant, indicating that the **fundamental negative information eventually dominates**, with vocal calm merely postponing, not preventing, the market's full adjustment.

Overall Insight: Voice as a Credibility Filter

Across all four quadrants, the market appears to use vocal cues as a **credibility filter**:

- **Aligned signals** (positive-positive or negative-negative) are taken at face value—optimism yields only modest lift, while genuine-sounding caution yields sharp underperformance.
- **Divergent signals** prompt investors to differentiate between strategic messaging and revealed conviction.

The results demonstrate that **executive tone meaningfully shapes return dynamics over 5–30 day horizons**, providing investors with information that is **absent from the transcript alone**. Vocal delivery adds a measurable layer of insight into conviction, comfort, and uncertainty, one that the market incorporates systematically into price behavior.

Conclusion

Speech Craft Analytics provides a new class of investable information derived from the *delivery* of corporate communication. While markets have digested text sentiment for more than a decade, the acoustic channel remains underexploited, despite being where involuntary signals of confidence, uncertainty, and stress naturally appear. Because executives can carefully manage what they say but have limited ability to manage *how* they say it, voice-based signals capture internal states that textual analysis cannot.

The empirical results in this white paper demonstrate that delivery-based factors generate persistent and economically meaningful excess returns across industries, market caps, and time horizons. Signals such as Nervousness, Balanced Delivery, PCA_AUDIO, and Uncertainty_mgmt exhibit strong monotonic decile spreads and robust performance under standard controls. These findings are not artifacts of modeling choices; they reflect a consistent behavioral mechanism in which the market underreacts to prosodic cues that reveal management's true assessment of risk, opportunity, and clarity of outlook.

For institutional investors, SCA offers a production-ready dataset that captures these latent behavioral dynamics at scale. Every earnings call is processed consistently, at the sentence level, producing speaker-specific features and model-derived factors aligned to real-time expectations. These signals can be incorporated into existing alpha frameworks exactly like any other numeric feed—ranked, bucketed, combined, or blended with text-based and market-based signals.

Investors seeking new, orthogonal sources of information, particularly those already fully optimized on traditional textual sentiment and fundamentals, now have access to a behavioral dimension the market has yet to price in. Voice is the final unmined modality in corporate disclosure, and the evidence suggests it contains durable and actionable information about future returns. SCA provides a turnkey way to capture that information and operationalize it in systematic strategies.

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

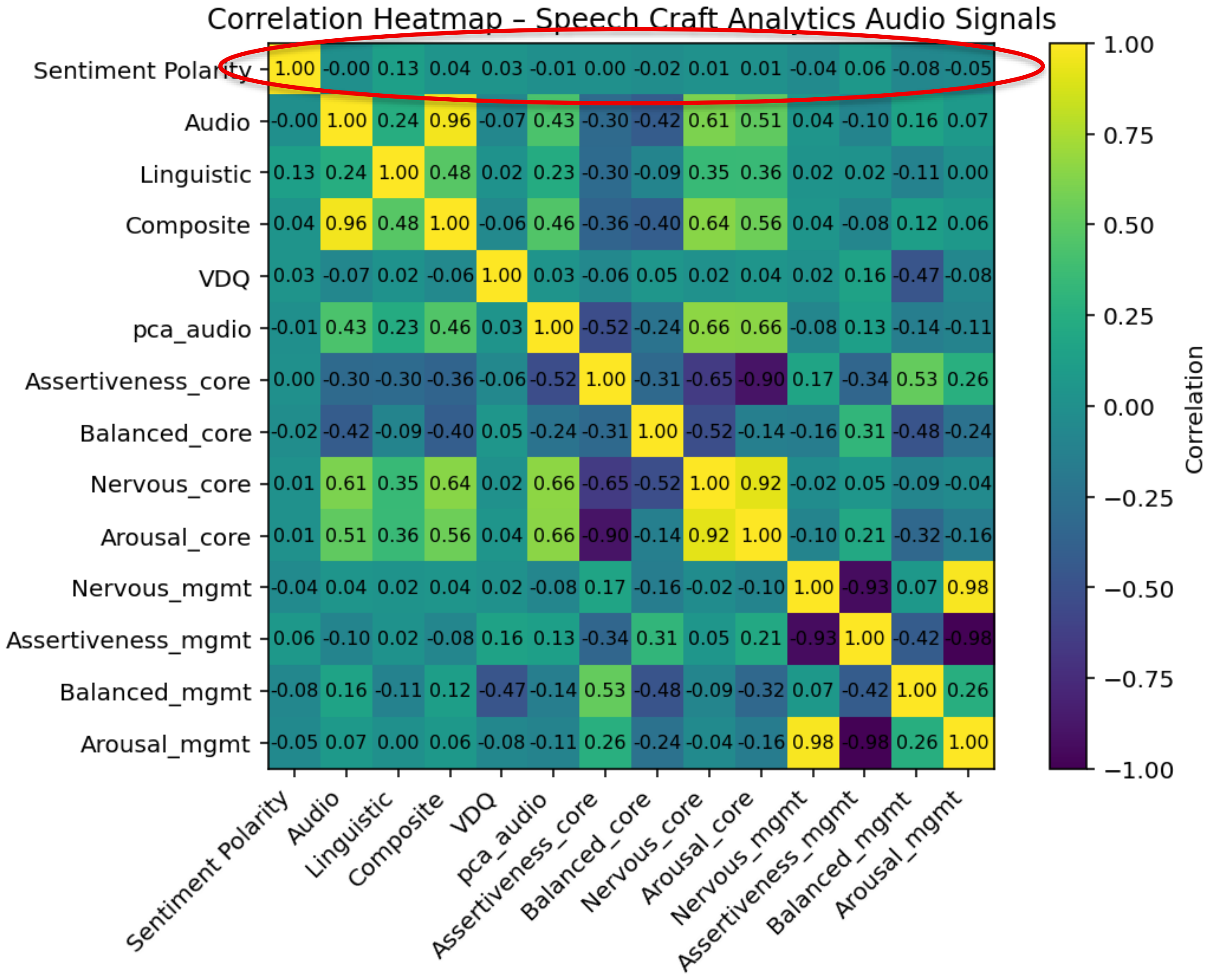
Wolfe Research (Luo's QES). (2023, June 14). The Leader's Voice: From transcripts to speech analysis in management presentations and conference calls. Wolfe Research LLC (industry report).

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

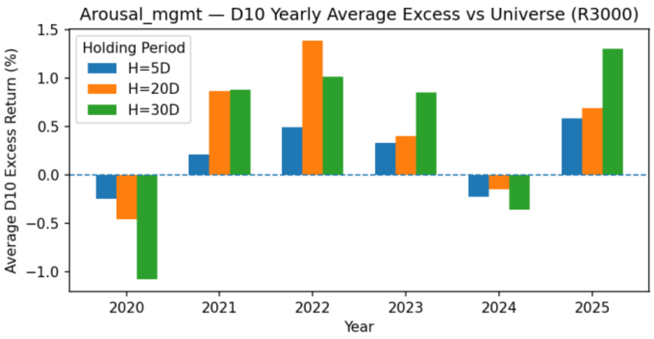
Speech Craft Analytics does not provide investment advice. Any information provided by Speech Craft Analytics, including data, analysis, reports, or any other communication, is solely for informational purposes and should not be considered a recommendation to invest, trade or engage in any financial 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.

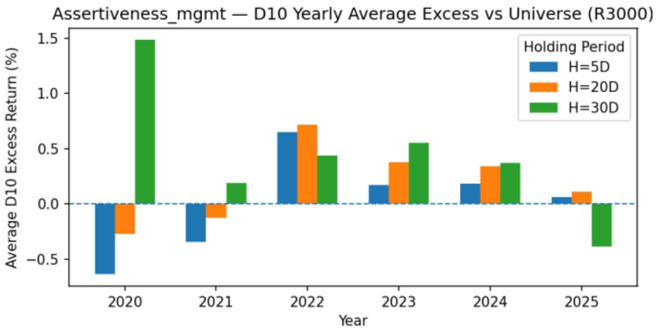
Appendix

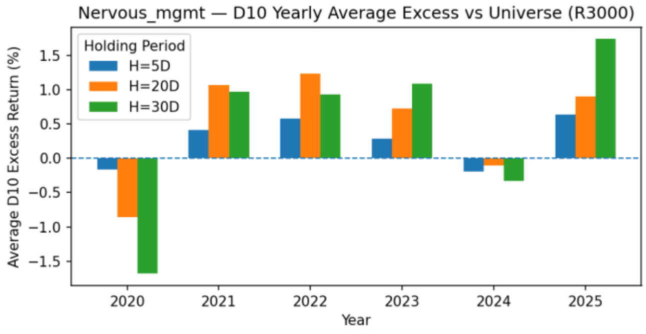
1. Correlation Matrix of SCA Audio Signals + Sentiment Polarity

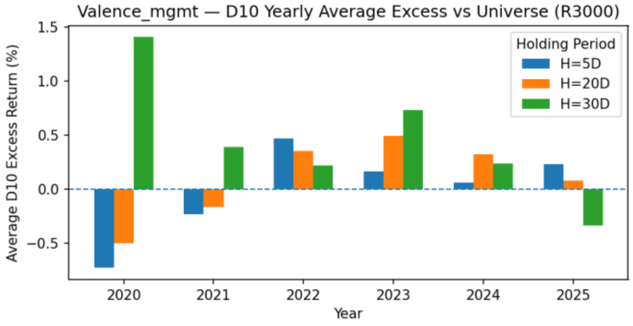


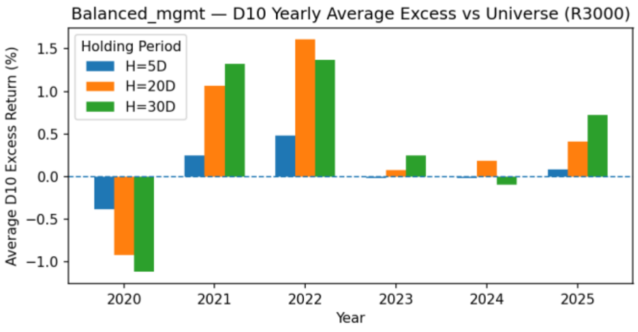
2. Yearly D10 Holding Period Average Excess Return

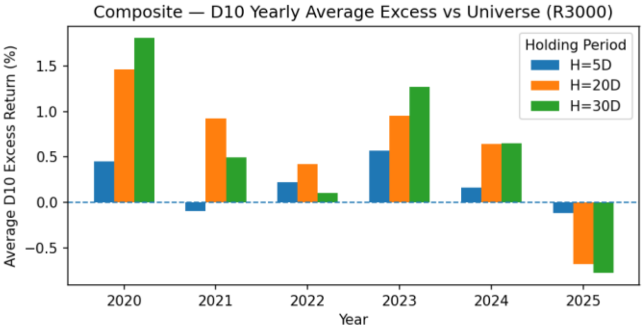


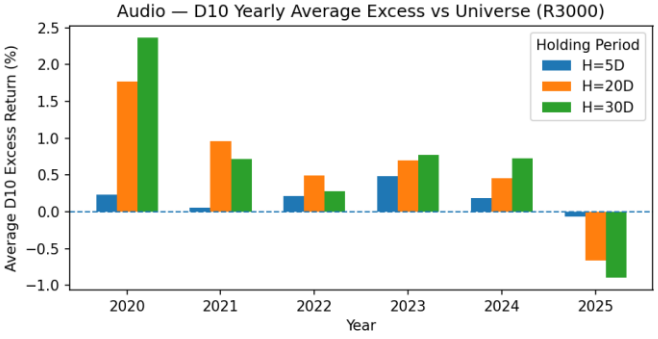












3. Case Study – Neutral Text but Nervous Management Tone

Comment by Anat Ashkenazi CFO Alphabet 2/4/25

"And for the um question with regards to where do we see or where do I see leverage moving forward and some of the comments I've made on the ah on the previous call, I certainly see opportunities for further productivity and efficiency, and this is one of ah our priority areas."

Listen Here

SCA Audio Score: -1.64 (Higher more confident)

SCA Composite: -4.06 (Higher more confident)

Probability Nervous: 98.9%

Sentiment Polarity: .02 (neutral)

Google Performance 2/5/25-3/7/2025: -9.08%

S&P 500 2/5/25-3/7/2026: = -4.7%

- **Tone Reveals What Text Conceals:**
  In responding to an analyst's question on operational efficiencies, a central concern as investors monitor funding for Alphabet's $75B CapEx, CFO Ashkenazi used reassuring language ("I certainly see opportunities…"). However, the **vocal delivery** told a different story. Slower speech, lowered pitch, and repeated phrasing signaled heightened nervousness even though the wording itself appeared neutral. There is even a voice waver on the work 'call'.

- **Investor Implication:**
  The neutral text paired with stressed vocal tone suggests **uncertainty beneath the surface** about Alphabet's ability to deliver the cost efficiencies being signaled. For investors, this tonal cue provides context not available from the transcript alone.

- **AI Highlights Hidden Tension:**

Speech Craft's AI detected the vocal stress, slowed cadence, pitch drop, and repetition, despite confident phrasing. These tonal markers indicate internal tension about executing large-scale efficiency initiatives.