What NLP Sentiment Can't Hear - *How Vocal Delivery Improves Earnings-Call NLP Sentiment*

June 2026

Abstract

This research brief examines whether managerial vocal delivery improves the economic interpretation of earnings-call text sentiment. Building on evidence that paralinguistic features of executive speech contain economically relevant information, we analyze 41,395 Russell 3000 earnings-call observations from July 15, 2020 through September 30, 2025, covering 2,862 unique tickers. Rather than asking whether voice predicts returns when text is neutral, this study examines whether vocal delivery helps investors interpret explicitly positive and negative textual sentiment. We sort earnings-call observations by NLP sentiment and management-calibrated vocal measures, then evaluate event-sample-relative excess returns over 10-, 20-, and 30-trading-day horizons. The findings show that vocal delivery adds incremental information to text sentiment, with the strongest evidence on the downside. In the lowest NLP sentiment quintile, negative text delivered with high Weak Constructive Delivery underperforms negative text delivered with low Weak Constructive Delivery by approximately 103 basis points over 20 trading days, with an event-date clustered t-statistic of -3.13. Low Balanced Delivery produces similar downside separation. Positive text also benefits from a voice layer: in the highest NLP sentiment quintile, positive text paired with high vocal Valence generates an additional 39 basis points over 10 trading days relative to positive text alone, with a clustered t-statistic of 2.03. The results suggest that vocal delivery is not a substitute for NLP sentiment, but an incremental conditioning layer that helps distinguish routine positive or negative language from language delivered with confirming vocal tone, controlled delivery, or weaker constructive vocal support.

Key Takeaways

Voice improves positive NLP sentiment when delivery confirms the words.
In the highest NLP sentiment quintile, positive text paired with high vocal Valence produces stronger 10-trading-day returns and higher hit rates than positive text alone.

Voice sharply conditions negative NLP sentiment.
The same negative-text bucket produces materially different return profiles depending on whether management delivery is constructive and controlled or weaker and less controlled.

Voice is a conditioning layer, not a substitute for NLP.
Text identifies the direction of management language; voice helps assess whether that language is delivered with confirming vocal tone, controlled delivery, or weaker constructive vocal support.

1. Introduction

Earnings-call NLP sentiment has become a standard way to convert management language into structured investment data. Text is scalable, sentiment can be measured consistently, and positive, negative, and neutral language can be ranked across companies, quarters, and speakers. This approach builds on a large literature showing that corporate language contains economically relevant information and that domain-specific textual analysis can improve interpretation of financial disclosures (Loughran and McDonald 2011; Larcker and Zakolyukina 2012).

But earnings calls are not ordinary documents. They are live communication events. Management is not only choosing words; management is delivering those words under pressure. NLP sentiment can identify favorable or unfavorable language, but it cannot hear whether positive language is delivered with vocal confirmation or whether negative language is delivered with weaker constructive vocal support. Prior research shows that nonverbal and paralinguistic features of executive speech can contain information about managerial affect, reporting credibility, and future firm outcomes (Mayew and Venkatachalam 2012; Hobson, Mayew, and Venkatachalam 2012).

This creates a practical limitation for text-only sentiment models. Positive language delivered with strong vocal Valence may carry a different signal than positive language delivered flatly or with weaker vocal confirmation. Negative language delivered with control may not carry the same implication as negative language delivered with high Weak Constructive Delivery. In both cases, the words alone are incomplete: text identifies the direction of the language, while voice helps investors assess confirmation, control, or weaker constructive vocal support.

A prior SSRN brief, *Voice Beyond Words: Evidence That Managerial Tone Predicts Returns When Text Does Not*, examined cases where textual sentiment was neutral and found that voice contained economically relevant information when NLP had little directional signal. This brief asks a different question: when NLP sentiment is already positive or negative, does vocal delivery make that sentiment signal more useful? The evidence suggests that it does. High vocal Valence strengthens positive NLP

sentiment, while Weak Constructive Delivery and reduced Balanced Delivery make negative NLP sentiment materially more concerning.

2. Data and Methodology

2.1 Data

The audio files for this analysis are sourced from S&P Global and cover available Russell 3000 earnings calls from July 15, 2020 through September 30, 2025. The primary H=20 analysis includes 41,395 earnings-call observations across 2,862 unique tickers; eligible counts are similar at the H=10 and H=30 horizons, with small differences due to return-window availability. The analysis focuses on the question-and-answer portion of the calls, where management responses are less scripted and more likely to reflect real-time communication under investor questioning. To focus on senior management communication, the sample is restricted to remarks by the Chief Executive Officer and Chief Financial Officer, or equivalent top-ranked management speakers.

Speech Craft Analytics performs its own transcription of the calls in order to preserve features of spoken communication that are often removed from commercial transcripts, including filler words, repetitions, repairs, and disfluencies. These elements are important because vocal delivery and spoken-language structure are part of the communication event. To reduce noise from short or incomplete utterances, observations with fewer than four alpha tokens are excluded.

2.2 Text and Voice Measures

Text sentiment is measured using Sentiment Polarity, a continuous measure of positive or negative textual tone. For the main tests, Sentiment Polarity is converted into percentile ranks and sorted into quintiles. The lowest quintile represents the most negative text, while the highest quintile represents the most positive text. This approach reflects the broader financial-text literature showing that language in corporate disclosures and conference calls can contain information relevant to investors (Loughran and McDonald 2011; Larcker and Zakolyukina 2012).

Voice measures are proprietary Speech Craft Analytics measures generated from management-calibrated vocal models. These models summarize executive vocal delivery along economically interpretable dimensions, including nervousness, entropy, arousal, assertiveness, vocal Valence, and Balanced Delivery. The use of management-calibrated vocal measures is motivated by evidence that vocal information in earnings calls can be economically meaningful, as well as by recent work emphasizing the importance of measuring vocal tone in corporate disclosure settings rather than relying only on generic speech-emotion constructs (Mayew and Venkatachalam 2012; Hobson, Mayew, and Venkatachalam 2012; Ewertz 2025; Pope 2026b).

The analysis emphasizes three voice constructs.

First, **Weak Constructive Delivery** is defined as:

$$\text{Weak Constructive Delivery} = \text{Nervousness} + \text{Entropy} + \text{Arousal} - \text{Assertiveness} - \text{Valence} - \text{Balanced Delivery}.$$

Weak Constructive Delivery is intended to capture a broad delivery state in which constructive vocal support is reduced. It increases when management exhibits higher nervousness, entropy, or arousal, and decreases when management exhibits higher assertiveness, vocal Valence, or Balanced Delivery.

Second, **Balanced Delivery** captures whether management's vocal delivery is more controlled, measured, and stable. Balanced Delivery is used as an independent check on the Weak Constructive Delivery result. If negative text is more concerning when delivered under strain, it should also be more concerning when delivered with lower Balanced Delivery.

Third, **vocal Valence** captures the positive or negative affective quality of the speaker's vocal delivery. In the positive-text tests, vocal Valence is used to examine whether favorable language becomes more informative when it is delivered with a more positive vocal tone.

All voice measures are SCA features that are expressed as percentile ranks and sorted into quintiles.

2.3 Experimental Design

This brief extends prior evidence that voice contains information when text sentiment is neutral. The earlier neutral-text design asks whether voice matters when NLP has little directional signal. The current design asks a different question: whether voice improves the interpretation of explicitly positive and negative NLP sentiment.

The analysis proceeds in three steps.

First, observations are sorted into quintiles based on Sentiment Polarity. The primary negative-text tests focus on the lowest text sentiment quintile, while the primary positive-text tests focus on the highest text sentiment quintile. Broader Q1–Q2 and Q4–Q5 text buckets are also examined as supporting tests.

Second, within each text bucket, observations are sorted by vocal delivery. For negative text, the main comparisons are high versus low Weak Constructive Delivery and low versus high Balanced Delivery. For positive text, the main comparison examines whether positive text paired with high vocal Valence improves on positive text alone. Supporting positive-text tests also examine high vocal Arousal and high Positive Affect.

Third, returns are compared within the same text sentiment bucket. This is the key identification idea: the text category is held constant, so differences in subsequent returns reflect the incremental information associated with vocal delivery.

The primary negative-text comparison is:

$$(\text{Negative NLP Text Sentiment} + \text{High Weak Constructive Delivery}) - (\text{Negative NLP Text Sentiment} + \text{Low Weak Constructive Delivery}).$$

A negative spread indicates that strained delivery makes negative text more concerning. The corresponding Balanced Delivery comparison is:

$$(\text{Negative NLP Text Sentiment} + \text{Low Balanced Delivery}) - (\text{Negative NLP Text Sentiment} + \text{High Balanced Delivery}).$$

Again, a negative spread indicates that weaker or less controlled delivery makes negative text more concerning.

The primary positive-text comparison is:

$$(\text{Positive NLP Text Sentiment} + \text{High Vocal Valence}) - (\text{Positive NLP Text Sentiment Alone}).$$

A positive spread indicates that favorable language becomes more informative when it is delivered with confirming positive vocal tone.

3. Results

The main evidence in this brief is strongest on the downside, but positive text also benefits from a voice layer. The relevant question is not only whether high-Valence positive statements outperform low-Valence positive statements. **The cleaner question is whether adding voice improves the signal already produced by positive NLP sentiment.**

For each holding period $H \in \{10,20,30\}$, the analysis uses winsorized H-day excess returns. The exhibit values are reported as **event-sample-relative excess returns**, defined as the observation's winsorized excess return minus the average winsorized excess return across all eligible observations at the same horizon:

$$\text{Event-Sample-Relative Excess}_{i,H} = \text{Excess}_{i,H} - \bar{\text{Excess}}_{H}.$$

This transformation centers each horizon around the average earnings-call event in the analyzed sample. Positive values indicate performance above the average event at that horizon; negative values indicate performance below the average event.

T-statistics are clustered by event date to account for common event-day effects across companies reporting on the same day. For matrix cells, t-statistics test each cell mean against zero. For contrast tables, t-statistics test the difference in means between the signal and comparison groups. Counts, hit rates, mean event-sample-relative excess returns, spreads, and clustered t-statistics are reported for the main exhibits.

Exhibit 1. Same Negative Text, Different Voice, Different Return

The 5×5 matrix below sorts observations by NLP sentiment quintile and Weak Constructive Delivery quintile. The key test is within the most negative text bucket, where text sentiment is held constant and vocal delivery varies.

Weak Constructive Delivery quintile	Q1 Low Text	Q2	Q3 Mid Text	Q4	Q5 High Text
Q5 High Weak Constructive Delivery	-0.67% (-2.63)	0.11% (0.46)	0.12% (0.46)	0.01% (0.05)	0.25% (1.08)

Weak Constructive Delivery quintile	Q1 Low Text	Q2	Q3 Mid Text	Q4	Q5 High Text
Q4	0.16% (0.57)	0.34% (1.34)	-0.21% (-0.88)	-0.15% (-0.66)	0.00% (0.01)
Q3 Mid Weak Constructive Delivery	-0.59% (-2.49)	-0.07% (-0.29)	-0.35% (-1.50)	0.08% (0.33)	0.23% (0.91)
Q2	0.43% (1.86)	0.06% (0.23)	-0.10% (-0.42)	-0.11% (-0.45)	-0.31% (-1.28)
Q1 Low Weak Constructive Delivery	0.36% (1.44)	0.11% (0.49)	-0.20% (-0.84)	-0.03% (-0.12)	0.41% (1.52)

Note: Cells show H=20 event-sample-relative excess return, with event-date clustered t-statistics in parentheses.

Within the lowest NLP sentiment quintile, high Weak Constructive Delivery observations produce an H=20 event-sample-relative excess return of -0.67%, while low Weak Constructive Delivery observations produce +0.36%. The same text bucket therefore has sharply different return outcomes depending on vocal delivery. The pattern should not be interpreted as a perfectly monotonic dose-response across all strain quintiles. Rather, the exhibit shows that negative-text observations are materially weaker when delivery is strained or less clearly controlled, with the pre-specified high-versus-low contrast producing the clearest separation.Exhibit 2. Voice-Conditioned Negative Sentiment Across Horizons

Horizon	NLP condition	Voice condition	Controlled-delivery return	Strained / weaker-delivery return	Spread	t-stat	Hit-rate change
H=10	Negative text Q1	High Weak Constructive Delivery vs Low Weak Constructive Delivery	+0.18%	-0.32%	-0.50%	-1.88	-4.35 pts
H=20	Negative text Q1	High Weak Constructive Delivery vs Low Weak Constructive Delivery	+0.36%	-0.67%	-1.03%	-3.13	-4.39 pts
H=30	Negative text Q1	High Weak Constructive Delivery vs Low Weak Constructive Delivery	+0.15%	-0.87%	-1.01%	-2.46	-2.11 pts

The pattern is consistent across horizons. The spread is negative at H=10, H=20, and H=30. It is economically meaningful at H=10, strongest at H=20, and remains large at H=30. This reduces the likelihood that the result is a one-horizon artifact.

Exhibit 3. Balanced Delivery Confirms the Same Interpretation

Horizon	NLP condition	Voice condition	Controlled-delivery return	Weaker-delivery return	Spread	t-stat
H=10	Negative text Q1	Low Balanced Delivery vs High Balanced Delivery	+0.14%	-0.27%	-0.41%	-1.57
H=20	Negative text Q1	Low Balanced Delivery vs High Balanced Delivery	+0.36%	-0.52%	-0.89%	-2.73
H=30	Negative text Q1	Low Balanced Delivery vs High Balanced Delivery	+0.23%	-0.73%	-0.97%	-2.24

Balanced Delivery provides a separate test of the same idea. Negative text is more concerning when delivered with less controlled vocal delivery. The signal is not simply “nervousness.” It is a broader delivery pattern: strain, uncertainty, reduced balance, and weaker control.

Exhibit 4. Convergence and Divergence

Regime	Definition	Interpretation
Positive convergence	Positive text + constructive voice	Words and delivery align positively
Positive divergence	Positive text + strained voice	Favorable words, less reassuring delivery
Negative convergence	Negative text + strained voice	Negative words and strained delivery align
Negative divergence	Negative text + constructive voice	Negative words delivered with greater control

Horizon	Contrast	Spread	t-stat
H=10	Negative divergence minus negative convergence	+0.46%	1.74
H=20	Negative divergence minus negative convergence	+0.76%	2.25
H=30	Negative divergence minus negative convergence	+0.72%	1.68

This is the convergence/divergence argument in its most useful form. The same negative text sentiment can mean different things depending on the voice. Negative language delivered under strain is not equivalent to negative language delivered with control.

The positive-text results provide a complementary test of the same conditioning-layer argument. The answer is yes, particularly at the 10-trading-day horizon. In the highest NLP sentiment quintile, positive-text observations produced an H=10 event-sample-

relative excess return of +0.07%. When those same positive-text observations were paired with high vocal Valence, the return increased to +0.46%. The incremental return was +39 basis points, with an event-date clustered t-statistic of 2.03. The hit rate also improved from 49.0% for positive text alone to 51.7% when positive text was paired with high vocal Valence.

Horizon	Positive-text signal	Positive text alone	Positive text + high Valence	Incremental return	t-stat	Hit-rate lift
H=10	Q5 NLP sentiment + Q5 vocal Valence	+0.07%	+0.46%	**+0.39%**	**2.03**	+2.72 pts
H=20	Q5 NLP sentiment + Q5 vocal Valence	+0.12%	+0.46%	**+0.34%**	**1.43**	+1.83 pts
H=30	Q5 NLP sentiment + Q5 vocal Valence	+0.24%	+0.73%	**+0.48%**	**1.65**	+2.14 pts

Note: *Cells report event-sample-relative excess returns. The incremental return compares the high-Valence overlay with the full positive-text Q5 benchmark at the same horizon. T-statistics are event-date clustered.*

The interpretation is that voice can help distinguish routine positive language from positive language delivered with stronger vocal confirmation. Positive earnings-call language is often expected, polished, and promotional. A text-only model may identify favorable wording, but it cannot determine whether the delivery sounds consistent with that favorable framing. Vocal Valence adds that layer.

The result is still asymmetric. Positive text paired with high vocal Valence improves on positive text alone, especially at H=10. But the more robust and economically sharper finding remains on the downside: negative text is much more strongly conditioned by Weak Constructive Delivery and reduced Balanced Delivery. Positive vocal Valence helps confirm favorable language; Weak Constructive Delivery helps identify when unfavorable language sounds materially worse.

4. Audio Illustration: American Express — Positive Text, Weak Vocal Confirmation

The positive-text results are more modest in the broad sample, but the concept is easy to hear in individual examples. American Express provides a useful illustration because the transcript and the voice point in different directions.

In one earnings-call exchange, management described consumer activity with the phrase "…spending continues to be strong." A conventional NLP sentiment model would naturally classify that language as positive. The words are favorable, direct, and easy to score as constructive.

The voice layer asks a different question: does the delivery provide the same positive confirmation as the words? In this example, the vocal signal provided weaker constructive support than the favorable wording suggested. That makes the sentence an example of positive-text / weaker-vocal-confirmation divergence.

Audio example: American Express management makes a strong positive statement with Weak Constructive Delivery. [Listen to the clip.]

The purpose of the example is not to make an investment claim from one sentence. It is to make the measurement problem tangible. Text identifies what was said. Voice helps determine whether the delivery confirms the words, weakens them, or raises a question that text alone would miss.

This example also clarifies the paper's asymmetric empirical finding. Positive-text divergence is intuitive and audible in specific cases, but the broad-sample evidence is strongest on the downside: when management uses negative language, voice provides a more powerful separation between controlled disclosure and disclosure delivered with weaker constructive vocal support.

5. Limitations and Interpretation

The results should be interpreted as evidence that vocal delivery conditions the return implications of NLP sentiment, not as proof that voice reveals managerial intent or deception. Weak Constructive Delivery is a measured delivery state, not a psychological diagnosis. The economic interpretation is that when negative language is delivered with Weak Constructive Delivery or reduced Balanced Delivery, the market appears to process that communication differently over subsequent trading days.

The analysis also shows that voice does not improve NLP sentiment symmetrically. Positive text shows some evidence of confirmation through vocal affect, but the strongest and most stable results appear in negative-text observations. This asymmetry is consistent with this view of earnings calls: positive language is often expected and polished, while negative language requires more contextual interpretation.

6. Voice as a Conditioning Layer Across Sentiment Regimes

Our earlier SSRN brief, *Voice Beyond Words: Evidence That Managerial Tone Predicts Returns When Text Does Not*, showed that voice matters when text sentiment is neutral; when NLP has little directional signal. This brief extends that evidence in a different

direction. The neutral-text finding says: when the words do not say much, voice can still carry information. **The current finding says: when the words are positive, voice helps determine whether the language is delivered with confirming vocal tone; when the words are negative, voice helps determine how concerning they are.**

Together, the two results define a coherent role for voice in earnings-call analysis. Voice is most useful when text is incomplete, ambiguous, or in need of interpretation. That includes neutral text, where NLP is muted; positive text, where vocal Valence helps distinguish routine favorable language from more convincing positive communication; and negative text, where Weak Constructive Delivery and reduced Balanced Delivery help distinguish controlled disclosure from strained disclosure.

7. Discussion

For investors already using NLP sentiment, the issue is not whether text contains information. It does. The issue is signal quality inside the same text bucket. A positive NLP score does not tell the investor whether favorable language is delivered with confirming vocal tone. A negative NLP score does not tell the investor whether unfavorable language is delivered with controlled delivery or weaker constructive vocal support. Voice adds a second-stage filter that helps re-rank NLP sentiment signals by delivery quality.

In the highest NLP sentiment quintile, positive text paired with high vocal Valence produces stronger subsequent returns than positive text alone. That result suggests that voice can help identify which favorable language deserves more weight. Positive text is common on earnings calls; positive text delivered with confirming vocal tone is more selective.

The downside result is stronger. In the lowest NLP sentiment quintile, negative text paired with high Weak Constructive Delivery produces materially weaker subsequent returns than negative text delivered with stronger constructive vocal support. Negative text paired with low Balanced Delivery shows a similar pattern. For an investor, this means negative NLP sentiment should not be treated as one uniform signal. Some negative language is routine caution. Some is delivered with control. Some is delivered with weaker constructive vocal support. Voice helps separate those cases.

This is the practical value of voice/text convergence and divergence. Convergence helps identify when the words and delivery point in the same direction: favorable words delivered with positive vocal Valence, or unfavorable words delivered with weak constructive vocal support. Divergence helps identify when the words and delivery do not carry the same implication: favorable words delivered with weaker vocal confirmation, or unfavorable words delivered with greater control.

The practical use case is not voice instead of NLP; it is voice on top of NLP. Voice gives investors a way to re-rank NLP sentiment signals after the text model has done its job. NLP sentiment identifies the direction of management language. Vocal delivery helps determine how much weight to put on that language.

8. Conclusion

NLP sentiment captures what management said. Voice helps interpret how management said it. The evidence in this brief shows that combining text and vocal delivery produces a more decision-useful signal than text alone.

The positive-text evidence shows that voice can confirm favorable language. In the highest NLP sentiment quintile, high vocal Valence improves the H=10 event-sample-relative excess return from +0.07% for positive text alone to +0.46%, with a clustered t-statistic of 2.03 and a hit-rate improvement from 49.0% to 51.7%.

Stronger evidence appears on the downside. In the lowest NLP sentiment quintile, high Weak Constructive Delivery underperforms low Weak Constructive Delivery by approximately 50 basis points over 10 trading days, 103 basis points over 20 trading days, and 101 basis points over 30 trading days. Low Balanced Delivery produces similar downside separation.

The implication is direct: NLP sentiment identifies the direction of the words; vocal delivery helps determine whether those words are reinforced by confirming vocal tone, supported by controlled delivery, or weakened by reduced constructive vocal support. Voice does not replace text. It makes text-based sentiment more investable.

References

Ewertz, S. (2025). Listen Closely: Measuring Vocal Tone in Corporate Disclosures, Journal of Accounting Research. August 2025.

Hobson, J. L., W. J. Mayew, and M. Venkatachalam. 2012. “Analyzing Speech to Detect Financial Misreporting.” *Journal of Accounting Research* 50, 349–392.

Larcker, D. F., and A. A. Zakolyukina. 2012. “Detecting Deceptive Discussions in Conference Calls.” *Journal of Accounting Research* 50, 495–540.

Loughran, T., and B. McDonald. 2011. “When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks.” *Journal of Finance* 66, 35–65.

Mayew, W. J., and M. Venkatachalam. 2012. “The Power of Voice: Managerial Affective States and Future Firm Performance.” *Journal of Finance* 67, 1–43.

Petersen, M. A. 2009. “Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches.” *Review of Financial Studies* 22, 435–480.

Pope, D. 2026. “Voice Beyond Words: Evidence That Managerial Tone Predicts Returns When Text Does Not.” SSRN working paper.

Pope, D. 2026. “Vocal Delivery as a Novel Risk Indicator: Evidence from Corporate Earnings Calls.” *Journal of Portfolio Management*.