

Hawkish by Voice

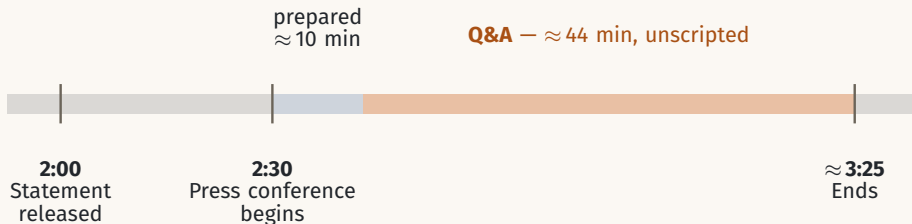
Financial Market Reactions to FOMC Communication

Yueh-Heng Wu · Hui-Ching Chuang · Ju-Fang Yen

National Taipei University

Research Seminar · 2026

The transcript is public within the hour. The market still watches for fifty-four minutes.

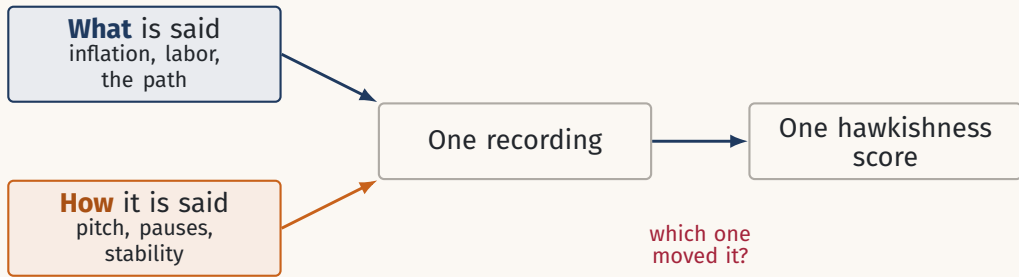


By 2:00 the decision is known. By 2:35 the words are transcribed and machine-readable.

Traders keep the feed on anyway — for the part nobody wrote down in advance.

■ If only the words carried information, the last fifty-four minutes would be redundant. Empirically they are not.

Content and delivery arrive together, so neither can be read off the other



A passage sounds hawkish. Is that because the words invoke inflation pressure — or because the delivery made otherwise neutral language sound restrictive?

| Scoring the original audio alone cannot answer this. The two channels are perfectly confounded in every recording ever made.

A small gap, priced in Treasuries, state-dependent in equities

1. Holding the words *exactly* fixed, original Q&A speech reads as **more hawkish** than a synthetic rendition of the same sentences. Small — **0.012** at the meeting level — but systematic.
2. A larger gap predicts **higher** two-day Treasury CARs across the curve: **8, 19, 32, 63** basis points from 1–3Y out to 20Y+. The conventional path surprise in the same regression goes the *other* way.
3. Equities and volatility respond **only when uncertainty is high**. At high EPU: SPY +57 bp, VIX –1.1, VVIX –3.1. Unconditionally, nothing.

| The rest of this talk is about whether you should believe finding 1, because findings 2 and 3 depend entirely on it.

| Three questions

1. For *identical* sentence content, does hawkishness inferred from original speech differ from hawkishness inferred from standardized synthetic speech?

A measurement question. Sections 1–4.

2. Is the meeting-level gap associated with market reactions, conditional on textual stance, policy surprises, and the macro state — and does it depend on policy uncertainty?

An asset-pricing question. Section 5.

3. Which observable acoustic dimensions does the speech model actually respond to?

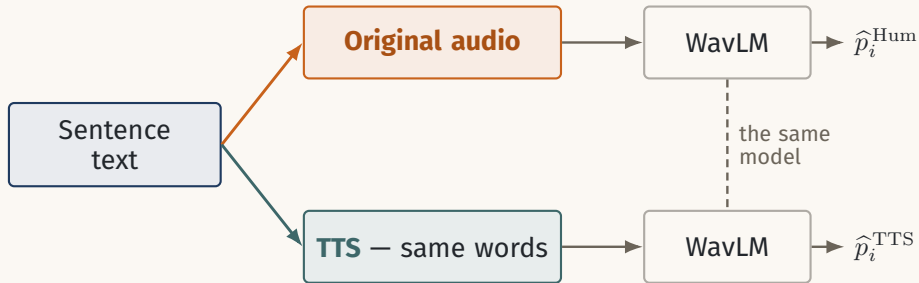
An interpretation question. Section 6.

| Question 1 is a measurement question. Questions 2 and 3 are only worth asking if the answer to 1 is yes.

The Fixed-Text Contrast

How to hold the words still

Every sentence gets a synthetic twin with identical words



$$\text{gap}_i = \hat{p}_i^{\text{Hum}} - \hat{p}_i^{\text{TTS}}$$

■ The TTS input is the ASR text, unedited. Rewriting even one word would destroy the property the whole design rests on.

Differencing removes the verbal contribution by construction, not by assumption

One model f_θ , two inputs that share every word:

$$\hat{p}_i^{\text{Hum}} = f_\theta(\mathbf{x}_i^{\text{Hum}}), \quad \hat{p}_i^{\text{TTS}} = f_\theta(\mathbf{x}_i^{\text{TTS}}), \quad \underbrace{\text{text}_i}_{\text{held fixed}}$$

Compare the usual alternative — score human audio, put a text control on the right:

- That needs the control *correctly specified*. Ours does not.
- That leaves residual content variation. Ours has none, by construction.
- The price: the benchmark voice is a **choice**, not a fact of nature.

■ This buys a within-sentence comparison, not exogenous variation. It is a measurement design, not an identification strategy — and we will not call it one.

This measures the Q&A's delivery — not the chair, and not causally

- **Not a chair effect.** The Q&A baseline pools chair answers, reporter questions, and moderator remarks. Apparent regime differences may be *composition*, not Powell's larynx.
- **Not a counterfactual.** One synthetic voice per chair regime cannot speaker-match a room of reporters of different genders and vocal characteristics. It is a standardized benchmark.
- **Not causal.** ΔS_m is not randomly assigned. Everything that follows is a conditional association, with 89 observations and year-clustered standard errors.

These limits are stated here, before the results, rather than at the end. They constrain what every coefficient later in this talk is allowed to mean.

Building the Sample

From 89 recordings to 35,619 scored sentences

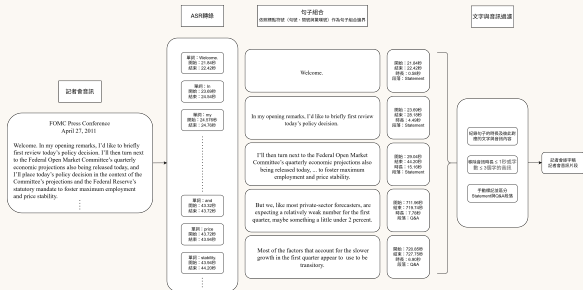
89 conferences, three chairs, fifteen years — and the panel is badly unbalanced



Press conferences began in April 2011 and have followed every scheduled meeting since 2019. Two-thirds of the sample is one chair.

Chair-regime dummies are in every regression. With 12 Bernanke conferences they are a control, not a comparison we can defend.

Word-level timestamps are what let a sentence carry its own audio



Whisper large-v3 transcribes; stable-ts stabilizes the word boundaries. Sentences — not speaking turns — are the unit: a turn mixes several stances, and a five-minute answer has no single delivery.

Without word-level alignment there is no way to score the delivery of a *specific* sentence, and the fixed-text design is impossible.

Filtering costs 17% of sentences and buys usable acoustics

Stage	Count	Why it changes
Raw conferences	89	2011–2025, three chairs
ASR + segmentation	42,819	Whisper large-v3, sentence assembly
Cleaning	35,619	Drop clips ≤ 1 s and ≤ 3 words
<i>of which</i> Statement	4,948	Prepared, scripted
<i>of which</i> Q&A	30,671	Interactive — the estimation sample
High-confidence binary	8,256	Text label hawkish/dovish at $p \geq 0.80$
Training slices	8,687	20-s windows, 2-s overlap

Slicing expands computational inputs, not statistical observations: window probabilities are averaged back to one number per sentence.

“Thank you.” is a real sentence with real audio and no monetary-policy stance. Dropping it is the difference between measuring delivery and measuring turn-taking.

FOMC-RoBERTa supervises the speech model — it is not the measure

We need labels for 8,000-odd audio clips. Hand-labeling is costly and inconsistent, so a text classifier labels the **words** and we transfer the label to the **audio**.

- The target is monetary-policy **stance**, not sentiment. A calm, upbeat delivery is not thereby dovish — which is where affect-trained classifiers go wrong.
- Keep only sentences labeled hawkish or dovish at $p \geq 0.80$: **35,619** $\rightarrow$ **8,256**, nearly balanced (4,209 / 4,047).
- Neutral and low-confidence sentences are dropped from *fine-tuning*, retained for *inference*.

█ Pseudo-label noise attenuates the speech model. It does not manufacture a human-TTS gap, because both arms inherit exactly the same label.

Learning Stance from Sound

WavLM, and why 8,256 labels are enough

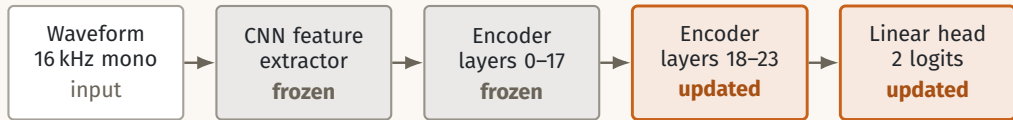
Self-supervised pretraining is what makes 8,256 labels enough

Training a speech classifier from scratch on 8,256 clips would fail. It does not have to be trained from scratch.

- WavLM was pretrained on $\approx 94,000$ hours of *unlabeled* speech by predicting masked audio, with simulated overlapping speakers and noise.
- That pretraining already encodes phonetics, prosody, speaker traits and channel conditions. None of it required a single label.
- Our 8,256 labels do one narrow job: point an existing representation at hawkish-versus-dovish.

■ This is transfer learning doing the work that our sample size cannot. It is also why partial freezing — next frame — is not a detail.

| Most of the model is inherited; we fit only the top quarter



≈ 317M parameters in total. We update roughly a quarter of the encoder.

| Everything to the left of the amber boundary is inherited from 94,000 hours of speech. Everything to the right is fitted on 89 press conferences.

Lower layers hear phonetics, upper layers carry meaning — so we freeze eighteen

- Layer-wise probing of self-supervised speech models shows an autoencoder-like profile: acoustic and phonetic information concentrates in early-to-middle layers, word-adjacent information higher up (Pasad et al., 2021; Chen et al., 2022).
- Fine-tuning all 24 layers on 8,256 *pseudo*-labeled clips would fit the labeler's noise, not the stance.
- So: freeze the CNN and layers 0–17; update layers 18–23 and the head.

This is a bias–variance choice, and it is worth being explicit that it was made *a priori* from the layer-wise literature — not selected on validation AUC.

Had we tuned the freezing depth against the validation folds, the 0.9125 we are about to report would be an in-sample number wearing a costume.

| The classifier is a two-logit head over a frozen backbone

For audio segment s with waveform $\mathbf{x}_s$, mean-pool the final hidden states into $f_\theta(\mathbf{x}_s)$ and read off two logits:

$$\mathbf{z}_s = \mathbf{W}f_\theta(\mathbf{x}_s) + \mathbf{b} = (z_s^H, z_s^D)'$$

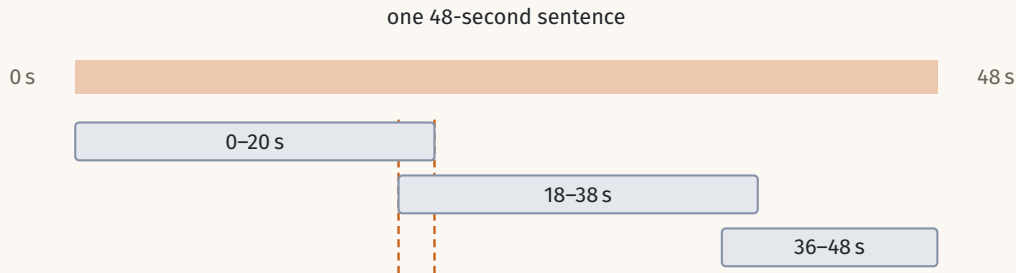
The model-implied hawkish probability is the binary softmax:

$$\hat{p}_s^H = \frac{\exp(z_s^H)}{\exp(z_s^H) + \exp(z_s^D)}$$

We keep $\hat{p}_s^H$ continuous rather than thresholding it here, preserving variation in how strongly the model reads a clip.

| Note there is no neutral class. The head is binary because the training screen kept only confidently directional sentences.

Long sentences become overlapping twenty-second windows



Two-second overlap, eighteen-second stride; only 4.4% of high-confidence sentences exceed 20 s. Window probabilities are averaged back to the sentence,

$$\hat{p}_i = |W_i|^{-1} \sum_{s \in W_i} \hat{p}_s^H.$$

■ Slicing changes the number of computational inputs, never the number of statistical observations. The unit stays the sentence.

Grouping the folds by conference is the difference between 0.91 and a leak

Sentences from one press conference share the microphone, the room, the chair, the policy context, and the day's news. A random split puts siblings on both sides of the validation line.

- **Grouped:** every sentence and every window from a given conference lands in the same fold. `StratifiedGroupKFold(groups=meeting)`.
- **Stratified:** the hawkish–dovish balance is preserved across folds.
- The model is therefore validated on *conferences it has never heard*, which is the only version of the number worth reporting.

■ An ungrouped split here would not be a small optimism. It would let the model recognize the day rather than the stance.

Out-of-fold AUC is 0.913, and symmetric across stances

Stratified group 5-fold, grouped by conference

Fold	AUC	Accuracy	Macro F1
1	0.9209	0.8495	0.8491
2	0.9076	0.8297	0.8293
3	0.9065	0.8363	0.8349
4	0.9065	0.8317	0.8283
5	0.9174	0.8350	0.8349
Pooled	0.9125	0.8362	0.8362

Out-of-fold performance by stance

Class	Prec.	Recall	F1
Dovish	0.8336	0.8350	0.8343
Hawkish	0.8387	0.8374	0.8380

Fold AUCs span a 0.014 band, so no single conference split carries the result.

Precision and recall differ by less than a percentage point between hawkish and dovish.

■ The model is not achieving accuracy by leaning on one class. That matters because our regressor is a *difference* of hawkish and dovish counts.

| Standardized voices, generated once per chair regime

Regime	Period	Microsoft neural voice
Bernanke	2011-04 – 2014-01	en-US-ChristopherNeural
Yellen	2014-03 – 2018-01	en-US-JennyNeural
Powell	2018-03 – 2025-12	en-US-ChristopherNeural

Rate, volume and pitch stay at their **defaults**. Tuning them per sentence would reintroduce exactly the delivery variation the benchmark exists to remove.

Voice is held fixed *within* a regime so the benchmark does not drift mid-sample. It is not speaker-matched: the Q&A pools reporters and a moderator of varying gender and vocal character, and one synthetic voice cannot stand in for all of them.

| This is the design's softest joint, and the most obvious place for a referee to push.

Does Delivery Register?

The fixed-text comparison, four ways

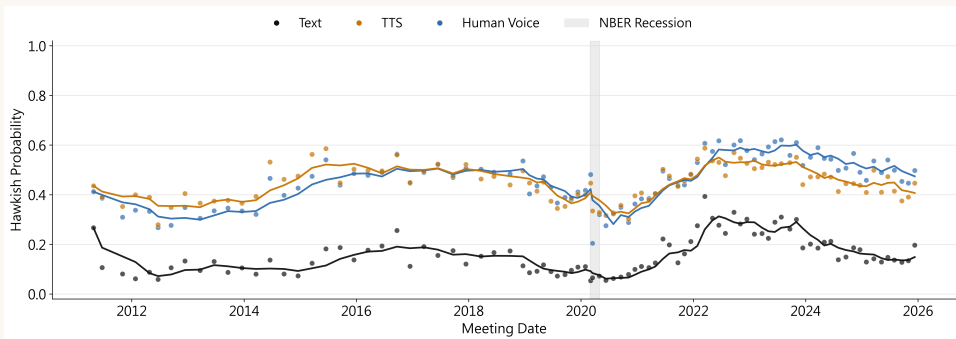
Human speech reads as more hawkish than its own transcript, spoken by a machine

Level	Section	<i>N</i>	Human	TTS	Gap	<i>t</i>
Sentence	Statement	4,948	0.5597	0.5378	0.0218	10.21***
Meeting	Statement	89	0.5714	0.5556	0.0158	2.68**
Sentence	Q&A	30,671	0.4225	0.4125	0.0100	9.94***
Meeting	Q&A	89	0.4589	0.4469	0.0120	2.24*

Positive and significant at every level and in both sections. The meeting-level gap survives duration-weighted aggregation, so it is not a handful of odd sentences.

■ The Q&A gap is 0.0100 on a probability scale — small. What follows depends on whether something that small can still be priced.

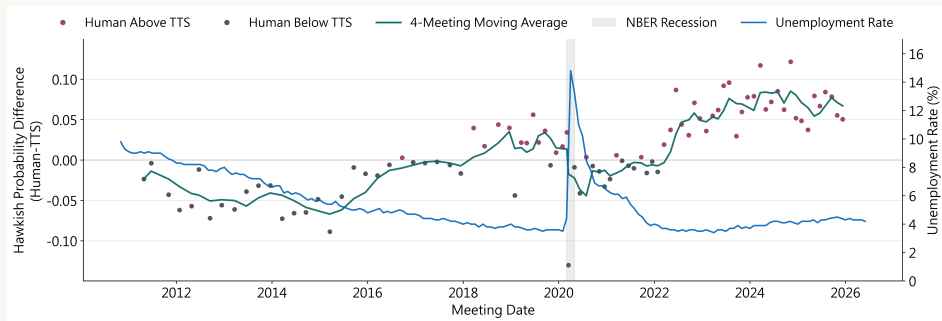
All three measures track the policy cycle



Low through 2011–2014, rising over the 2015–2018 normalization, collapsing in early 2020, rising again after 2022.

Text, TTS, and human hawkishness all know what decade it is. That is a sanity check on the speech model, not a finding.

The gap changes sign — it is not a constant machine-versus-human offset



Predominantly negative before 2016, crossing around 2018, unsettled through the pandemic, consistently positive after 2022.

■ A fixed synthesis artifact would be a constant. This moves with the meeting, which is the necessary condition for it to explain anything.

The gap is audible: wider pitch range, more shimmer, less loudness

Feature	Sentence	Meeting	Reading
F0 Mean	0.3448**	0.7217	pitch level, similar
F0 SD	-1.1863***	-0.8150	TTS varies pitch more
F0 Range	35.1062***	45.0890***	humans swing far wider
Loudness Mean	-7.1236***	-8.0002***	TTS is louder
Loudness SD	-7.9843***	-6.3064***	TTS is more variable
Jitter	-0.0021***	-0.0025***	TTS jitters more
Shimmer	0.0217***	0.0182***	humans waver in amplitude
HNR	-1.6422***	-1.1801***	TTS is cleaner, less noisy

Eight Praat features, human minus TTS, Q&A. Six of the eight survive duration-weighted aggregation to the meeting level; the two pitch-level measures do not.

■ The gap is not an artifact of WavLM's output layer. The raw acoustics differ in the same paired design, without any classifier involved.

The synthetic series is a flat line; the human series is a cloud



Eight meeting-level acoustic features, 2011–2025. **TTS** in orange, **human** in blue.

Read the *shape*, not the panels:

- Orange sits in a **flat band** in every panel — the benchmark is standardized by construction.
- Blue **scatters** in all eight — delivery moves meeting to meeting.
- The pink stripe is the Yellen regime, where the synthetic voice switches. Orange jumps level; blue does not follow it.

That last point matters: the human series is not tracking the synthesizer. If it were, the gap would be an artifact of our voice choice rather than a fact about the Q&A.

Do Markets Price It?

Ten series, 89 events, two days

Abnormal returns are measured against a sixty-day pre-event mean

$$CAR_{m,a,[0,h]} = \sum_{\ell=0}^h AR_{m,a,\ell}, \quad AR_{m,a,\ell} = r_{m,a,\ell} - \bar{r}_a^{\text{pre}}$$

- Ten series: SPY; SHY, IEI, IEF, TLT; TIP; HYG, LQD; VIX, VVIX.
- ETFs enter as abnormal *log returns*; VIX and VVIX are indices, so they enter as abnormal *level changes*.
- Main window $[0, 1]$; robustness runs $[0, 0]$ through $[0, 15]$.
- The statement lands at 2:00 and the conference starts at 2:30 — the event measured is the conference, not the decision.

■ A constant-mean model, not a market model. With 89 events, adding estimated betas would cost more in noise than it removes in bias.

High-confidence counts, not average probabilities, define net hawkishness

$$S_m^{(k)}(\tau) = \frac{N_{m,H}^{(k)}(\tau) - N_{m,D}^{(k)}(\tau)}{N_{m,H}^{(k)}(\tau) + N_{m,D}^{(k)}(\tau)}, \quad k \in \{\text{Text, TTS, Hum}\}$$

Non-directional sentences leave the denominator entirely. The regressor is the gap:

$$\Delta S_m^{H-T}(\tau) = S_m^{(\text{Hum})}(\tau) - S_m^{(\text{TTS})}(\tau), \quad \tau = 0.80$$

Why counts and not means: a conference holds a few hundred directional sentences and several thousand near-0.5 ones. Averaging raw probabilities lets the indifferent majority drown the signal.

■ ΔS is bounded in $[-2, 2]$, realized in $[-0.329, 0.342]$, $\text{SD} = 0.158$. Every magnitude later is scaled by that SD, never by a unit change.

Twelve regressors, 89 observations, standard errors clustered by year

$$\begin{aligned} CAR_{m,a,[0,h]} = & \alpha + \beta_1 \Delta S_m^{H-T}(\tau) + \beta_2 S_m^{(\text{Text})}(\tau) + \gamma_1 MP1_m + \gamma_2 MP2_m \\ & + \eta_1 Unemp_m + \eta_2 PCE_m + \theta_1 EPU_m + \theta_2 (\Delta S_m^{H-T} \times EPU_m) \\ & + \delta' \mathbf{Z}_m + \varepsilon_{m,a,h} \end{aligned}$$

- $MP1, MP2$: policy-rate and expected-path surprises (Acosta et al. database).
- $\mathbf{Z}_m$: SEP indicator, Bernanke and Yellen dummies (Powell omitted).
- $Unemp, PCE, EPU$ mean-centered, so β_1 is read at average uncertainty.
- OLS, standard errors clustered by year.

■ Year clusters, not meeting-level HC1: conferences within a year share a macro regime. It is the conservative choice and it costs us precision.

The voice gap raises Treasury CARs, rising with maturity

	SHY [1–3Y]	IEI [3–7Y]	IEF [7–10Y]	TLT [20Y+]
Human–TTS	0.005* (1.915)	0.012** (2.052)	0.020** (2.301)	0.040*** (3.008)
Text	–0.003 (–0.856)	–0.005 (–0.456)	–0.004 (–0.252)	0.017 (0.678)
MP2 (path surprise)	–0.043*** (–10.835)	–0.095*** (–8.946)	–0.135*** (–5.670)	–0.150* (–1.702)
<i>N</i>	89	89	89	89
<i>R</i> ²	0.405	0.335	0.303	0.246

Scaled by one SD of the gap (0.158): **8, 19, 32, 63** basis points of two-day CAR. Textual stance is insignificant throughout — the text index is uniformly dovish (–0.97 to –0.46), so it carries little usable variation as a control.

Four assets, four positive coefficients, rising with duration. The next frame argues this is one finding, not four.

Scaled by each asset's own volatility, it is one result observed across the curve

	SHY	IEI	IEF	TLT
Basis points per SD of gap	8	19	32	63
In SD of the asset's own CAR	0.44	0.39	0.40	0.40

The maturity ordering is duration scaling, not four independent discoveries. Pairwise correlations among the four Treasury CARs run 0.547 to 0.966, and exceed 0.87 for adjacent maturities.

Reporting this as “significant in four of ten assets” would overstate the evidence. It is one Treasury result, seen four times.

The sign is the finding: the voice gap and the path surprise pull opposite ways

MP2 — expected-path surprise

A more restrictive path than expected
⇒ yields up, bond prices **down**.

-0.043, -0.095, -0.135, -0.150.

Standard duration logic. Unsurprising.

Human—TTS — the voice gap

Delivery read as *more* hawkish than the same text

⇒ bond prices **up**.

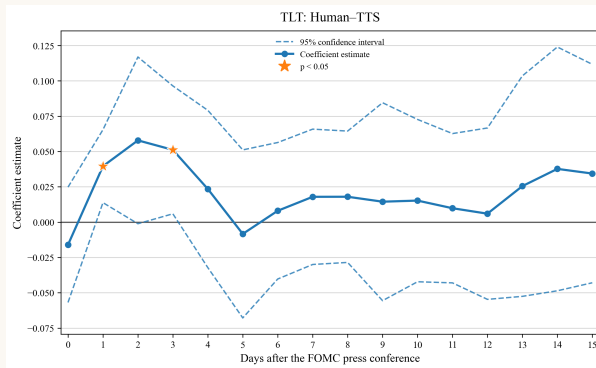
+0.005, +0.012, +0.020, +0.040.

Same regression, same day.

No expected-rate-path story produces both signs at once. Raw correlations agree: the gap correlates -0.085 with MP1 and -0.154 with MP2 — near-orthogonal before any conditioning.

| This is the strongest evidence that the voice gap is not a noisy restatement of a tightening surprise. It is also the finding we can least explain.

The response lives in the first three trading days



TLT coefficient traced across windows $[0, h]$, $h = 0, \dots, 15$. Precise at $h = 1$ and $h = 3$; intervals widen and cover zero thereafter.

Short-horizon repricing, not return predictability. A signal that still paid at day fifteen would be the more troubling result.

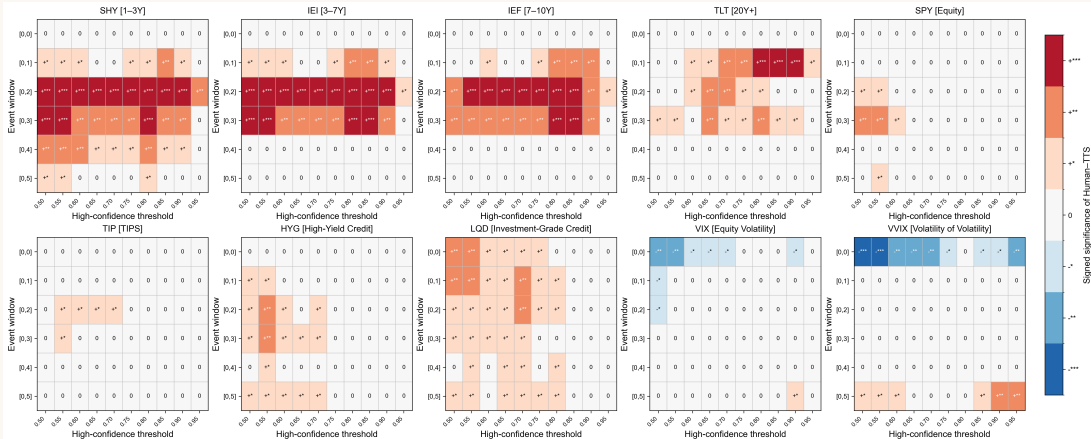
Equities and volatility respond only when uncertainty is high

	SPY [Equity]	TIP [TIPS]	HYG [HY]	LQD [IG]	VIX [Vol.]	VVIX [VoV]
Human—TTS	−0.007 (−0.269)	0.021 (1.443)	0.013 (0.803)	0.034* (1.868)	−4.057 (−0.798)	1.116 (0.109)
× EPU	0.0003*** (2.730)	0.000 (1.117)	0.000 (1.617)	0.000 (0.988)	−0.058*** (−3.135)	−0.163*** (−4.173)
MP2	−0.238*** (−2.912)	−0.138*** (−4.436)	−0.155*** (−4.398)	−0.185*** (−4.075)	13.218 (0.715)	13.170 (0.240)
<i>N</i>	89	89	89	89	89	89
<i>R</i> ²	0.195	0.280	0.206	0.362	0.252	0.170

Main effects are indistinguishable from zero for SPY, VIX and VVIX; the interactions are significant at 1%. At one SD of both the gap (0.158) and EPU (119.9): SPY +**57** bp, VIX −**1.1**, VVIX −**3.1** — about 0.3 SD of each dependent variable.

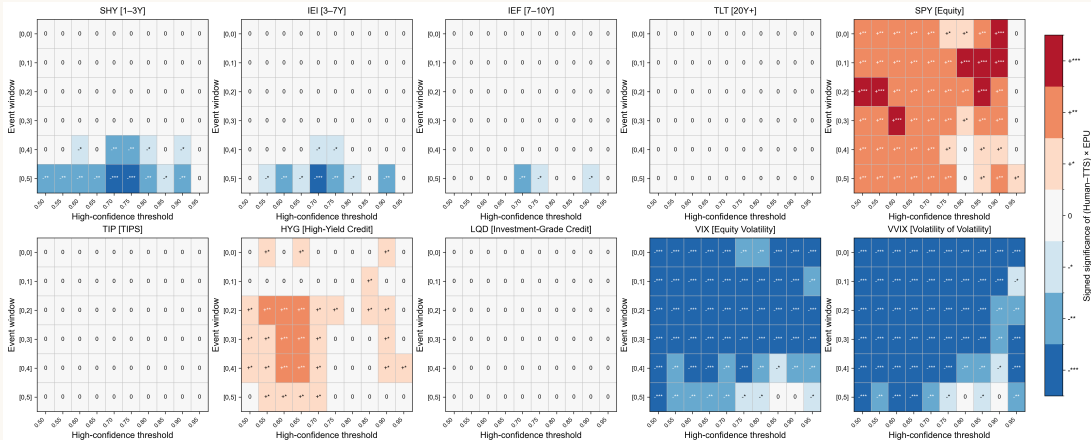
Equities rising while volatility falls is a risk-premium or clarity pattern. A pure tightening shock would push equities the other way.

Significance sits in contiguous regions, not isolated cells



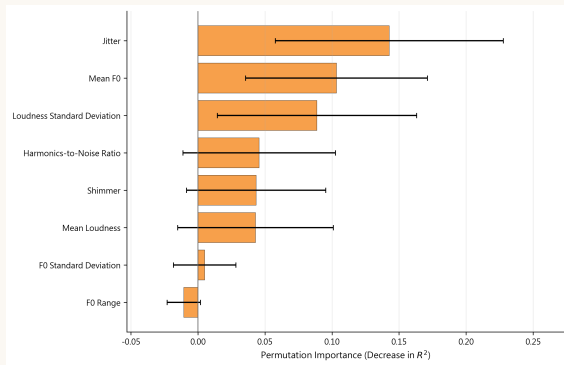
Read the blocks of colour, not the cells. We fixed $\tau = 0.80$ and $[0, 1]$ in advance — this panel shows what we would have found had we chosen otherwise.

The EPU interaction is systematic and opposite in sign



SPY positive across most thresholds and short windows; VIX and VVIX negative almost everywhere; Treasury interactions empty. Two coherent patterns, not a handful of cells.

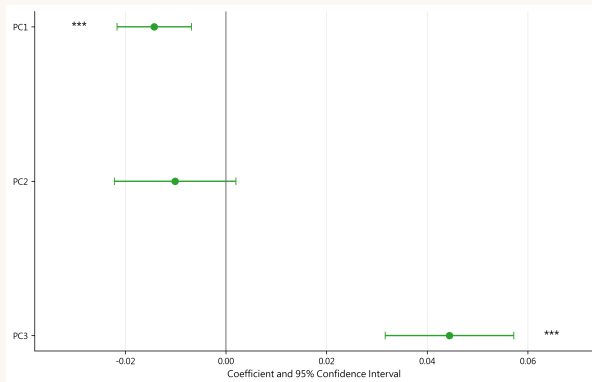
No single acoustic feature carries it — jitter merely leads



Permutation importance: mean drop in validation R^2 when one feature is shuffled, over 5 folds $\times$ 100 repetitions.

■ Error bars are wide and some cross zero. Read this as “distributed across perturbation, pitch and loudness”, not as a ranking.

Hawkish delivery is tense *and* controlled at once



PC1 = clarity vs. perturbation, enters *negatively*. **PC3** = vocal force and loudness control, enters *positively*. PC2: null.

More perturbation *and* more intensity control together — real-time pressure under a firm delivery. Descriptive: this is not a mediation test.

A Skeptic Would Say

Four objections we take seriously

Four objections, and what we can and cannot answer

Objection	Response
"This is just Powell's voice."	We do not claim it is. The measure pools reporters and the moderator, so composition shifts sit inside the variation. Conceded.
"The TTS voice is arbitrary — another one gives another gap."	True in level. The regressions use the <i>time-series</i> variation, and a fixed voice cannot generate that. Partly answered.
"89 observations, twelve regressors, clustered by year."	Correct — which is why we lean on the maturity pattern and the heatmap regions, not on any single <i>t</i> -statistic. Conceded.
"You searched over τ and h until something worked."	The heatmaps show every cell we could have reported. Significance forms contiguous regions in $[0, 2] - [0, 3]$. Answered.

Two of the four are conceded. The paper is evidence that delivery carries priced information, not evidence about whose delivery.

Four claims we can defend, and four we cannot

Supported

- Human and same-text TTS speech are scored differently, and differ acoustically.
- The gap moves with the meeting, not with the synthesizer.
- It predicts short-horizon Treasury CARs, distinct from MP1 and MP2.
- Equity and volatility effects appear only at high EPU.

Not established

- *Why* the sign is positive — credibility, clarity, term premia, composition all live.
- Whose delivery matters — chair, reporters, or the interaction.
- Any causal claim. ΔS_m is not assigned.
- Whether it survives a different TTS system.

■ The measurement result is solid. The economics of *why* is still open — and we would rather put that on a slide than round it off.

The sharpest untested implication: breakevens, not real rates

If the positive Treasury sign runs through **inflation expectations** — a Q&A perceived as more resolute raising confidence that the Committee will contain inflation — then the effect must appear in the *inflation-compensation* component of nominal yields, not the real component.

That is a sharp, falsifiable prediction:

- It should be visible in the **nominal-minus-TIPS breakeven** spread.
- If the effect sits in real rates instead, the credibility story is wrong and a term-premium or information story survives.

We do not run it here — it needs a breakeven design, not an ETF CAR.

■ We are telling you the test that could falsify our preferred interpretation, because we have not run it yet.

How the Fed *says* it carries information
the transcript does not.

Thank you

Hui-Ching Chuang

`hcchuang.github.io`

`huichingc@gmail.com`