Hawkish by Voice: Financial Market Reactions to FOMC Communication

Yueh-Heng Wu[a], Hui-Ching Chuang[a,*], Ju-Fang Yen[a]

[a]*National Taipei University, New Taipei City, Taiwan*

Abstract

This paper examines whether vocal delivery during Federal Open Market Committee (FOMC) press-conference Q&A sessions contains market-relevant information beyond the words spoken. Using WavLM, a large-scale speech processing model, we construct a vocal-based measure of monetary-policy stance, defined as hawkish versus dovish rather than generic emotion, applied to 89 press conferences from 2011 to 2025. For each sentence, we compare the original human audio with a text-to-speech (TTS) rendering of the same text. The Human–TTS difference provides a fixed-text measure of Q&A delivery and speaker composition. Larger gaps are associated with higher short-window cumulative abnormal returns on Treasury ETFs across maturities, conditional on textual stance, macroeconomic conditions, monetary-policy surprises, and chair regimes. When economic policy uncertainty is high, larger gaps are also associated with higher S&P 500 ETF returns and declines in VIX and VVIX. The evidence indicates that the vocal delivery and speaker composition of the Q&A contain state-dependent information not captured by text alone.

Keywords: Central bank communication, Vocal hawkishness, Text-to-speech perturbation, Financial-market reactions, Economic policy uncertainty, Self-supervised speech models
JEL: E52, E58, G12, G14, C45

*Corresponding author: Hui-Ching Chuang. Address: Department of Statistics, National Taipei University, No. 151, University Rd., Sanxia Dist., New Taipei City, Taiwan 237. Phone: 886-9727-35021. We thank Hsuan Fu and Hong-Wen Cheng for constructive comments and valuable insights. Hui-Ching Chuang and Ju-Fang Yen gratefully acknowledge financial support from the National Science and Technology Council, Taiwan.

Email addresses: ygsdsf111@gmail.com (Yueh-Heng Wu), hcchuang@gm.ntpu.edu.tw (Hui-Ching Chuang), jfyen@gm.ntpu.edu.tw (Ju-Fang Yen)

1. Introduction

The Federal Open Market Committee (FOMC) influences U.S. interest rates, asset prices, and financial conditions, and these influences are transmitted through financial markets. As central bank communication has grown in importance, market participants look beyond the policy-rate decision itself and read the statement, the Summary of Economic Projections, and the post-meeting press conference for signals about the future policy path. The press conference, which combines prepared opening remarks with an unscripted question-and-answer (Q&A) session, is a particularly rich source of such signals because it reveals the reasoning and stance behind the decision.

A press conference is not just a text document. The words exchanged during the Q&A convey judgments about inflation, the labor market, and the likely policy path; *how* participants deliver those words through pitch, loudness, pauses, speaking rate, voice stability, and changes in tone may also shape how markets interpret the message. In central bank communication, both the content ("what is said") and the delivery ("how it is said") hold importance, with the latter potentially conveying information that extends beyond the actual words. The challenge is that content and delivery occur simultaneously. When a passage sounds hawkish, this may be because the words invoke inflation pressure, the need for rate hikes, or a restrictive stance, or because its delivery makes otherwise neutral language sound more restrictive. Studying only the original human audio therefore cannot separate whether a model's inferred stance comes from the text or from delivery and speaker composition.

To address this measurement challenge, we introduce a TTS-based fixed-text benchmark. For every usable sentence in the press conference, we retain the sentence text, the original human audio, and a synthetic (TTS) rendering of the identical text. In the Q&A, the human audio includes the chair's answers, reporters' questions, and moderator remarks. Because each human clip and its TTS counterpart contain the same words and are scored by the same speech model, their difference holds verbal content fixed. We therefore interpret it as a delivery-related contrast for the full interactive Q&A, subject to the limitations of the selected TTS system and synthetic voices. We construct three sentence-level hawkishness measures: a text measure from FOMC-RoBERTa, a TTS measure from standardized synthetic speech, and a human measure from the original conference audio. Aggregating high-confidence sentence classifications over the full Q&A, we define the *human-minus-TTS net hawkishness gap* (the "voice gap") as human net hawkishness minus same-text TTS

net hawkishness.

The paper addresses three questions. First, for identical sentence content, does the hawkishness inferred from original Q&A speech differ systematically from that inferred from standardized TTS delivery? Second, is the meeting-level Q&A voice gap associated with post-conference market reactions after controlling for textual stance, monetary-policy surprises, and the macroeconomic environment, and does this association vary with economic policy uncertainty (EPU)? Third, which observable acoustic dimensions of the Q&A are associated with WavLM-implied hawkishness? These questions map to the fixed-text TTS comparison, the cumulative abnormal response (CAR) analysis, and the acoustic, random-forest, and principal-component analyses.

The sample comprises 89 post-meeting press conferences from April 27, 2011 to December 10, 2025, spanning the Bernanke, Yellen, and Powell chair regimes. We collect the full conference audio and transcripts, use automatic speech recognition (ASR) to construct sentence-level text–audio alignments, generate hawkish, dovish, and neutral text labels with FOMC-RoBERTa, and fine-tune WavLM on high-confidence hawkish and dovish examples. For the Q&A analysis, we retain all usable spoken sentences rather than restricting the sample to the chair's answers. We then score the original and matched TTS audio and relate meeting-level Q&A measures to market reactions and acoustic characteristics.

This paper contributes to the literature on central bank communication in three ways. Prior work has analyzed FOMC communication from several angles. One strand quantifies *policy surprises* and their market effects: Bernanke and Kuttner (2005) show that because asset prices are forward-looking, market reactions at announcements mainly reflect the unexpected component of policy, which they measure using federal funds futures. Gürkaynak et al. (2005) decompose announcements into a target-rate factor and a path factor, with the path factor exerting a larger influence on longer-maturity yields. Building on this framework, Swanson (2021) further separates federal funds rate, forward guidance, and asset-purchase shocks. Acosta et al. (2025) assemble a U.S. Monetary Policy Event-Study Database from which we draw the surprise measures MP1 and MP2 used as controls. A second strand studies the *information effect* and event-study design. Lucca and Moench (2015) document a pre-FOMC announcement drift in equities, and Pan and Peng (2024) document related drift in long-term Treasuries, underscoring the importance of event-window choice. Nakamura and Steinsson (2018) and Jarociński and Karadi (2020) show that

announcements reveal the central bank's assessment of fundamentals, so that market reactions reflect both policy and information shocks; relatedly, Aruoba and Drechsel (2024) apply natural language processing to internal Federal Reserve staff documents to capture the information set on which the Committee conditions when identifying policy shocks. A third strand *quantifies policy text*: Shah et al. (2023) label FOMC-related text as hawkish, dovish, or neutral—a framework underlying the FOMC-RoBERTa classifier we use—and Deng et al. (2024) integrate text and audio at fine granularity. Most directly related, Gorodnichenko et al. (2023) use a machine-learning measure of the chair's vocal tone in 36 FOMC press conferences from 2011 to 2019 and show that it contains information for equity returns beyond policy actions and textual sentiment. Complementing this audio evidence, Curti and Kazinnik (2023) use minute-level facial-expression measures from FOMC press-conference videos and find that negative expressions are associated with adverse market reactions after controlling for verbal content and other meeting characteristics. Together, these studies show that vocal and visual nonverbal channels in FOMC press conferences can carry market-relevant information.

Against this background, three gaps remain. First, existing voice studies mostly measure generic emotion or tone rather than hawkish–dovish policy stance, and they typically rely on classifiers trained on actor-recorded emotional speech; Gorodnichenko et al. (2023), for instance, train on the RAVDESS and TESS databases. Both features are limiting here. Hawkishness is a policy construct rather than an affective one, so a delivery that sounds calm or upbeat is not thereby dovish. And Ewertz et al. (2026) document that actor-trained models degrade to approximately chance accuracy on executives' unscripted speech in earnings conference calls—evidence from corporate disclosure rather than central banking, but press-conference Q&A exchanges are similarly spontaneous and uncontrolled. Second, they rarely construct a standardized TTS benchmark under *fixed* text, so content and delivery remain entangled. Third, the relationship between speech-model output and observable acoustic features is underexplored. Our contribution is threefold. We treat the FOMC press conference as a multimodal communication event that contains both text and voice; we construct a TTS-based fixed-text benchmark and define the human-minus-TTS voice gap as a delivery-related contrast for the full interactive Q&A under fixed content; and we examine whether the voice gap is related to market reactions, whether that relation is moderated by policy uncertainty, and what acoustic dimensions underlie the speech model's hawkishness,

using acoustic features, random forests, and principal component analysis (PCA). To construct the speech measures, we fine-tune WavLM (Chen et al., 2022), a self-supervised model that first learns transferable representations from large amounts of unlabeled speech, which makes it feasible to build a domain-specific hawkishness classifier from a limited set of high-confidence pseudo-labels (Pasad et al., 2021).

The main findings can be previewed as follows. First, holding verbal content fixed, the original full-Q&A speech is modestly but systematically more hawkish than its TTS rendering in the model-implied measure. The human and synthetic versions also differ along several acoustic dimensions, indicating that the fixed-text comparison captures delivery-related variation that is not fully reproduced by the selected TTS voices. Second, the market results are strongest for fixed-income assets. In the Q&A, a larger Human–TTS net hawkishness gap is associated with higher $[0, 1]$ cumulative abnormal returns for U.S. Treasury ETFs spanning short-, intermediate-, long-, and very-long-maturity securities, with weaker positive evidence for the investment-grade corporate-bond ETF, LQD. These estimates are distinct from the conventional expected short-rate-path surprise, which enters negatively and significantly across most fixed-income specifications. The voice gap therefore captures an additional dimension of monetary-policy communication rather than simply restating a standard tightening surprise.

Third, equity and volatility responses are state-dependent. The unconditional Human–TTS coefficient is not precisely estimated for SPY, VIX, or VVIX, but its interaction with economic policy uncertainty is economically and statistically important. When EPU is high, a larger voice gap is associated with higher SPY cumulative abnormal returns and with larger declines in VIX and VVIX. This pattern is more consistent with improved message clarity, greater credibility, or lower uncertainty and risk premia than with a pure tightening shock. Finally, the acoustic interpretation exercises show that WavLM-implied human hawkishness is related to a combination of vocal perturbation, pitch, and loudness control. Random-forest and principal-component analyses indicate that these associations are distributed across several acoustic dimensions rather than driven by a single feature. Taken together, the results suggest that the vocal delivery and speaker composition of the full Q&A contain incremental, state-dependent information beyond its verbal content.

The remainder of the paper is organized as follows. Section 2 describes the data, text and audio processing, and the fine-tuning of WavLM to classify hawkish and dovish monetary-policy

stance from speech. Section 3 introduces the fixed-text Human–TTS comparison and constructs the meeting-level vocal stance measures and acoustic analysis. Section 4 presents the empirical results. Section 5 concludes. Appendix Appendix A provides the WavLM fine-tuning details, and Appendix Appendix B characterizes the acoustic dimensions associated with WavLM-implied human hawkishness.

2. Data and Monetary Policy Stance Labeling

2.1. FOMC Press-Conference Data

Our sample comprises 89 post-meeting FOMC press conferences held between April 27, 2011, and December 10, 2025, spanning the chairmanships of Ben Bernanke, Janet Yellen, and Jerome Powell. Each conference contains prepared opening remarks and a media Q&A, with publicly available audio and transcripts that permit sentence-level text–audio alignment.[1] Within our sample, the conference dates are April 27, 2011–January 29, 2014 for Bernanke, March 19, 2014–January 31, 2018 for Yellen, and March 21, 2018–December 10, 2025 for Powell.

The FOMC typically holds eight scheduled meetings per year, approximately six weeks apart. These meetings generally take place on Tuesday and Wednesday, although the exact dates follow the Federal Reserve's official annual calendar. The Federal Reserve began holding post-meeting press conferences in April 2011 and has held one after every scheduled meeting since 2019. The policy decision and statement are typically released at 2:00 p.m. ET on the second meeting day, followed by the chair's press conference at 2:30 p.m. ET. The press conference therefore constitutes a distinct communication event occurring after markets have received the formal policy decision.

Each conference begins with prepared opening remarks averaging slightly less than ten minutes, followed by a moderated Q&A session lasting approximately 44 minutes. During the Q&A, the moderator calls on reporters in turn, who typically identify themselves and their media organizations before asking questions. Reporters mainly represent major financial and general news outlets, such as *The Wall Street Journal* and *Bloomberg*, and the chair answers an average of 22 questions and follow-ups per conference (De Pooter, 2021).

[1]Federal Open Market Committee, `https://www.federalreserve.gov/monetarypolicy/fomc_historical.htm`. Raw audio and related files are retrieved using the Python packages `requests` and `yt-dlp`.

The recorded Q&A consequently contains multiple speaker roles, including the chair's answers, reporters' questions, and occasional moderator remarks. Our baseline Q&A measure uses all usable sentences from this interactive segment. It should therefore be interpreted as capturing the vocal delivery and composition of the full Q&A, rather than the chair's voice alone.

Because the opening remarks are prepared in advance whereas the Q&A is interactive and less scripted, we use the Q&A as our primary empirical setting, as in Gorodnichenko et al. (2023). This choice follows conference-call research that treats the Q&A as a particularly informative communication environment (Borochin et al., 2018; Li et al., 2021; Ewertz et al., 2026). We depart from Gorodnichenko et al. (2023) in one respect: whereas they restrict their audio sample to the chair's answers, we retain all usable spoken sentences, so that the resulting measure captures the combined delivery of questions, answers, and moderation after markets have already received the formal policy decision and prepared remarks.

2.2. Automatic Speech Recognition, Sentence Construction

For each FOMC press conference, we collect the official transcript and the corresponding audio. The transcript provides an external reference for validating the automatic speech recognition (ASR) output and identifying the transition from the chair's prepared statement to the Q&A session.[2] Appendix Figure A.1 summarizes the construction of the sentence-level text–audio dataset. We first transcribe each press conference and obtain start and end timestamps for every recognized word.[3] These timestamps allow each textual unit to be matched to the precise segment of the original audio in which it was spoken.

We then combine recognized words into sentences using periods, question marks, and exclamation marks as sentence boundaries. For each sentence, we retain the ASR text, cleaned text, start and end times, audio duration, and word count, and extract the corresponding audio segment. We use sentences rather than complete speaking turns because they provide relatively coherent units of semantic content and vocal delivery, reduce the mixing of multiple monetary policy stances within

[2]We download the official transcript files from the Federal Reserve's website and extract the corresponding audio from the press-conference videos. These steps are automated using the Python `requests` library and `yt-dlp`, respectively. See `https://requests.readthedocs.io/en/latest/` and `https://github.com/yt-dlp/yt-dlp`.

[3]We use Whisper large-v3 (Radford et al., 2023) for English speech-to-text transcription and `stable-whisper`, also distributed as `stable-ts`, to improve timestamp stability and obtain word-level alignment. See `https://github.com/openai/whisper` and `https://github.com/jianfch/stable-ts`.

long utterances, and establish a common granularity for text labeling, speech-model training, and press-conference-level aggregation. This sentence-level design follows prior research that estimates textual signals at the sentence level (Huang et al., 2023) and work that aligns financial speech with transcripts to estimate vocal signals at comparable granularity (Deng et al., 2024; Alexopoulos et al., 2024; Ewertz et al., 2026).

We exclude audio segments lasting no more than one second and sentences containing three words or fewer. The minimum sentence-length restriction and the duration restriction provide a screen for clips containing insufficient acoustic information.[4] These filters reduce the initial 42,819 ASR sentences to 35,619 usable text–audio pairs, retaining 83.2% of the initial observations.

Finally, using the official transcript and the content surrounding the transition, we manually identify the boundary between the prepared statement and the Q&A session and assign each sentence to the corresponding section. We do not separate Q&A sentences by speaker role in the baseline design; reporter questions, chair responses, and moderator remarks are all retained. This distinction is important because the statement is prepared in advance and typically delivered in a structured manner, whereas the Q&A contains real-time exchanges, pauses, clarifications, and reformulations. The resulting section indicators allow all sentence-level measures and press-conference-level aggregates to be constructed separately for the prepared statement and the less-scripted Q&A session.

2.3. Monetary Policy Stance Labeling and Speech-Model Fine-Tuning

To construct a speech-based measure of monetary policy stance, we fine-tune WavLM-Large to distinguish hawkish from dovish speech. Fine-tuning requires sentence-level labels for the corresponding audio, but manually classifying tens of thousands of clips would be costly and could introduce inconsistency in the labeling standard. We therefore adopt a weak-supervision strategy in which FOMC-RoBERTa serves as an intermediate labeling model: it maps the verbal content of each sentence into a standardized monetary policy stance label, which is then assigned to the aligned audio.[5] FOMC-RoBERTa is therefore not our measure of vocal delivery; it provides scalable and

[4]Very short observations commonly consist of pleasantries or discourse markers, such as "Thank you," or arise from anomalous ASR segmentation. Such clips contain limited semantic content and are unlikely to provide a reliable measure of monetary policy stance or vocal delivery.

[5]We use `gtfintechlab/FOMC-RoBERTa`, released by the Georgia Tech Financial Services Innovation Lab on Hugging Face: `https://huggingface.co/gtfintechlab/FOMC-RoBERTa`. A pseudo-label is a model-generated

internally consistent supervision for adapting the speech model to the FOMC setting. Built on RoBERTa and adapted to FOMC language, it distinguishes hawkish and dovish monetary policy positions from generic positive and negative sentiment (Liu et al., 2019; Shah et al., 2023).

For each sentence, FOMC-RoBERTa produces hawkish, dovish, and neutral probabilities. We retain only sentences classified as hawkish or dovish with a maximum predicted probability of at least 0.80. This screen reduces noise from ambiguous pseudo-labels and yields a nearly balanced training sample of 8,256 sentences, comprising 4,209 hawkish and 4,047 dovish observations. Neutral and lower-confidence sentences are excluded from fine-tuning but retained for full-sample inference.

This design departs from the voice literature in two respects, both motivated in the introduction: the supervision target is monetary policy stance rather than generic emotion, and the adaptation audio is drawn from the FOMC setting itself rather than from the actor-recorded emotional-speech corpora on which affective classifiers are typically trained (Ewertz et al., 2026).

We fine-tune WavLM-Large, a self-supervised speech representation model pretrained on large corpora of unlabeled audio (Chen et al., 2022). All audio is standardized to 16,000 Hz mono waveforms. Clips longer than 20 seconds are divided into overlapping 20-second windows, with a two-second overlap between adjacent windows. Let $\mathbf{x}_s$ denote the waveform of audio window s, and let $f_\theta(\mathbf{x}_s)$ denote its WavLM representation. The classification layer produces the hawkish and dovish logits

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

The model-implied hawkish probability is

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

Rather than dichotomizing the model output, we retain $\widehat{p}_s^H$ as a continuous measure of speech-implied monetary policy stance, thereby preserving variation in the strength of the model's hawkish assessment. When a sentence is divided into multiple audio windows, we average the window-level probabilities to obtain a single sentence-level measure.

Because WavLM-Large is highly parameterized relative to the available FOMC training sample

training label rather than a manually assigned label.

and is fine-tuned using pseudo-labels, unrestricted updating may lead to overfitting. We therefore freeze the convolutional feature extractor and the first 18 Transformer encoder layers, updating only the final six layers and the classification head. This partial-freezing strategy preserves pretrained acoustic representations while allowing task-specific adaptation, consistent with evidence on layer-wise specialization in self-supervised speech models, documented for wav2vec 2.0 by Pasad et al. (2021) and reflected in the layer-wise evaluation of Chen et al. (2022). Appendix Appendix A provides further rationale and reports the complete regularization, optimization, and implementation settings.

2.4. Sample Construction and Model Inputs

Table 1 summarizes the data-processing pipeline. ASR transcription and sentence segmentation of the 89 FOMC press conferences produce 42,819 sentences. This initial sample includes short, non-substantive utterances, such as "Thanks." We remove audio clips lasting no more than one second and sentences containing no more than three words, leaving 35,619 valid sentence-level text–audio pairs. FOMC-RoBERTa classifies 25,423 sentences as neutral (71.37%), 5,043 as hawkish (14.16%), and 5,153 as dovish (14.47%). For WavLM fine-tuning, we retain only hawkish or dovish sentences with classification confidence of at least 0.80. This restriction yields 8,256 high-confidence sentences, comprising 4,209 hawkish and 4,047 dovish observations. Neutral and lower-confidence sentences are excluded from training but retained for full-sample inference.

Because our WavLM implementation accepts audio inputs of at most 20 seconds, longer sentences are divided into overlapping 20-second windows with a two-second overlap, equivalent to an 18-second stride. For example, a 48-second sentence is divided into three windows covering seconds 0–20, 18–38, and 36–48. The overlap provides a buffer around each boundary and reduces the loss of acoustic and linguistic context that would arise from non-overlapping segmentation. Only 4.43% of the high-confidence sentences exceed 20 seconds. Applying this rule increases the 8,256 training sentences to 8,687 training windows.

We apply the same windowing procedure when using the fine-tuned WavLM model to infer vocal hawkishness for all 35,619 cleaned sentences, producing 36,695 inference windows. When a sentence generates multiple windows, we compute the arithmetic mean of their predicted hawkish probabilities to recover one hawkishness measure for the complete sentence. Thus, windowing

expands the number of computational inputs but not the number of statistical observations. All subsequent sentence-level analyses, including the paired t-tests in Table 4, use the original complete-sentence unit. Accordingly, the 4,948 statement sentences and 30,671 Q&A sentences reported in Table 1 sum to the full sample of 35,619 sentences.

After constructing these inputs and fine-tuning WavLM-Large, we evaluate the resulting model using stratified group five-fold cross-validation, with press-conference date as the grouping variable. All sentences and audio windows from a given conference are assigned to the same fold, preventing observations from the same event which share recording conditions and policy context from entering both the training and validation samples. Stratification approximately preserves the hawkish–dovish class balance across folds.

Table 2 reports both aggregate and class-specific validation results. Fold-specific AUCs range from 0.9065 to 0.9209, and the pooled out-of-fold AUC is 0.9125; pooled Accuracy and Macro F1 are both 0.8362. Performance is also nearly symmetric across policy stances. The dovish class has Precision, Recall, and F1 of 0.8336, 0.8350, and 0.8343, respectively, while the corresponding values for the hawkish class are 0.8387, 0.8374, and 0.8380. The limited variation across folds and classes indicates that the model's predictive performance is not driven by a particular conference split or by systematically favoring one policy-stance class.

3. Human–TTS Fixed-Text Contrast

To isolate variation associated with vocal delivery, we generate a synthetic speech counterpart for each sentence. The TTS input is exactly the sentence text obtained from the ASR alignment procedure; it is neither rewritten nor otherwise modified. Each observation therefore forms a one-to-one match among the sentence text, the original human audio, and a TTS rendition of the same text. In the Q&A, the original audio may contain the chair, a reporter, or the moderator. Because the human and TTS recordings share identical words and syntax, their difference provides a fixed-text, delivery-related contrast. The TTS benchmark is not intended to reproduce any participant's individual voice, but to provide a standardized rendition of the same sentence.[6] We generate

[6]Microsoft describes neural text-to-speech as converting written text into human-like synthetic speech using deep neural networks. See `https://learn.microsoft.com/en-us/azure/ai-services/speech-service/text-to-speech`.

the synthetic audio using Microsoft neural voices. For consistency within each chair regime, we use `en-US-JennyNeural` during the Yellen period and `en-US-ChristopherNeural` during the Bernanke and Powell periods. Because the full Q&A includes reporters and moderators of different genders and vocal characteristics, this choice should be viewed as a standardized regime-specific benchmark rather than a speaker-matched counterfactual.[7]

For sentence i, we construct three hawkishness probabilities:

$$\widehat{p}_i^{(\text{Text})}, \qquad \widehat{p}_i^{(\text{TTS})}, \qquad \widehat{p}_i^{(\text{Hum})}, \tag{1}$$

where $\widehat{p}_i^{(\text{Text})}$ is the FOMC-RoBERTa hawkish probability estimated from the sentence text, $\widehat{p}_i^{(\text{TTS})}$ is the WavLM hawkish probability for the corresponding synthetic audio, and $\widehat{p}_i^{(\text{Hum})}$ is the WavLM hawkish probability for the original human audio of that sentence, which in the Q&A may be spoken by the chair, a reporter, or the moderator. When a sentence exceeds 20 seconds and is divided into multiple windows, we average its window-level predictions to obtain one sentence-level probability. The same WavLM model generates the human and TTS probabilities from audio containing identical verbal content. We therefore define the sentence-level human-minus-TTS gap as $\widehat{p}_i^{(\text{Hum})} - \widehat{p}_i^{(\text{TTS})}$. A positive value indicates that the original delivery of the sentence is assessed as more hawkish than its standardized TTS rendition. We interpret this difference as a delivery-related contrast rather than as a literal counterfactual treatment effect.

We construct meeting-level hawkishness probability, $H_m^{(k)}$, as the duration-weighted average of sentence-level hawkishness probabilities within press conference m, where $k \in \{\text{Text}, \text{TTS}, \text{Hum}\}$. The meeting-level Human–TTS gap is defined as $H_m^{(\text{Hum})} - H_m^{(\text{TTS})}$. The fixed-text design compares the original human speech with a TTS rendition of the identical sentence. Because the two recordings contain the same verbal content and are evaluated by the same speech model, their difference provides a delivery-related contrast. A positive meeting-level gap indicates that the original full-Q&A delivery is assessed as more hawkish than the corresponding same-text TTS benchmark.

Table 3 shows that the speech model assigns systematically higher hawkishness to human than to TTS audio in both the prepared statement and the Q&A. At the sentence level, the mean Human–TTS

[7]Synthetic audio is generated in batch using `edge-tts`, which provides access to Microsoft Edge online TTS voices and supports adjustments to voice, rate, volume, and pitch. See `https://github.com/rany2/edge-tts`.

gaps are 0.0218 and 0.0100, respectively, both significant at the 0.1% level. After duration-weighted aggregation, the meeting-level gaps remain positive at 0.0158 and 0.0120, significant at the 1% and 5% levels. The persistence of the difference at the meeting level indicates that it is not confined to isolated sentences.

The Q&A provides the more informative setting for the market analysis because it is interactive and less scripted than the prepared statement. Conditional on identical words, the full-Q&A gap can reflect real-time variation in emphasis, rhythm, pitch, vocal stability, turn-taking, and the changing composition of questions and answers. The statement results provide a benchmark showing that Human–TTS differences are also detectable in prepared communication. However, because the Q&A measure pools all speaker roles, it should not be interpreted as a chair-only delivery effect.

Figure 1 presents the time-series evidence. Panel A shows that text, TTS, and human hawkishness share the broad monetary policy cycle: hawkishness is relatively low during 2011–2014, rises during the 2015–2018 normalization, declines sharply in early 2020, and increases again during the post-2022 tightening cycle. This common movement indicates that all three measures retain information about the underlying policy environment. Panel B shows that the meeting-level Human–TTS gap changes sign and magnitude over time. It is predominantly negative early in the sample, becomes more positive around 2018, fluctuates during the onset of the pandemic, and is generally positive after 2022. The gap therefore captures meeting-specific variation in delivery rather than a constant difference between human and synthetic speech.

We next examine whether the model-implied hawkishness gap corresponds to observable acoustic differences. For each sentence, we extract the same eight acoustic features from the human and TTS audio and calculate the Human–TTS difference in each feature. We then aggregate these feature gaps from the sentence level to the meeting level separately for the statement and Q&A sections.

The eight features capture complementary dimensions of vocal delivery. Mean F0 measures the average pitch level; F0 STD measures within-clip pitch variation; and F0 Range is the difference between the maximum and minimum F0. Mean Loudness measures average vocal intensity, while Loudness STD captures within-clip variation in intensity. Jitter measures cycle-to-cycle irregularity in fundamental frequency, and Shimmer measures cycle-to-cycle irregularity in amplitude. HNR is the harmonics-to-noise ratio, with higher values indicating a clearer, more periodic, and more

stable voice signal. Appendix Appendix B details the acoustic features and uses random-forest and principal-component analyses to characterize the acoustic dimensions associated with WavLM-implied human hawkishness.

Table 4 reports paired mean differences defined as Human minus TTS. Human speech has a substantially wider F0 Range and greater shimmer than the corresponding TTS speech in both the statement and Q&A. Human speech also has a slightly higher mean F0 at the sentence level, although this difference is not statistically significant after meeting-level aggregation. By contrast, human speech has lower F0 variation, lower average loudness, lower loudness variation, lower jitter, and lower HNR than TTS speech. Most of these differences remain statistically significant at the meeting level. Thus, the fixed-text Human–TTS comparison reveals systematic differences in pitch, intensity, periodicity, and voice quality rather than variation arising solely from the WavLM classification output.

Figure 2 further shows that the TTS acoustic series is comparatively standardized (stable) across meetings, whereas the original full-Q&A speech displays substantially greater temporal variation. Because these meeting-level Q&A aggregates combine the voices of the chair, reporters, and moderator, apparent differences across chair regimes may also reflect changes in participant composition. They should not be interpreted as intrinsic differences in the chairs' vocal characteristics or policy stance.

Taken together, the paired probability tests, time-series patterns, and acoustic evidence support interpreting the Human–TTS gap as a fixed-text measure of full-Q&A delivery. By differencing WavLM outputs for identical text, the measure removes the component common to the original and synthetic renditions while retaining variation associated with delivery and speaker composition. It does not isolate the chair's individual vocal contribution.

4. Market Reactions

We examine market responses across ten financial series spanning equities, Treasury securities, inflation-protected bonds, corporate credit, and market volatility. Table 5 summarizes the instruments and daily response construction. SPY represents the U.S. large-cap equity market. SHY, IEI, IEF, and TLT capture Treasury securities with maturities of 1–3, 3–7, 7–10, and more than 20 years, respectively. TIP represents inflation-protected Treasuries, while HYG and LQD capture

below-investment-grade and investment-grade corporate bonds. For these eight ETFs, the daily response is the log return computed from adjusted closing prices.

We also examine the VIX and VVIX indices. VIX measures short-horizon implied volatility in the U.S. equity market, whereas VVIX measures the implied volatility of VIX options and thus captures uncertainty about future volatility. Because VIX and VVIX are indices rather than traded ETFs, their daily responses are measured as changes in index levels rather than returns. Daily observations are obtained from Yahoo! Finance and cover the 2011–2025 press-conference sample, including the pre-event estimation windows required to construct abnormal market responses.

Let t_m denote the event day for conference m. We use the 60 trading days preceding the event to estimate the asset's normal daily response. The abnormal response on relative event day ℓ is the observed response minus this pre-event mean, and the cumulative abnormal response over $[0, h]$ is

$$CAR_{m,a,[0,h]} = \sum_{\ell=0}^{h} AR_{m,a,\ell}. \tag{2}$$

For the eight ETFs, $CAR_{m,a,[0,h]}$ is a cumulative abnormal log return; for VIX and VVIX, it is a cumulative abnormal change in the index level. We use CAR as common notation for both constructions. The main results use the $[0, 1]$ window, while robustness tests consider windows from $[0, 0]$ through $[0, 15]$.

For the market analysis, we focus on directional policy signals rather than the full distribution of sentence-level probabilities. We therefore construct a high-confidence hawkishness index that summarizes the relative frequency of hawkish and dovish sentences within each press conference. We then difference the corresponding human-speech and TTS indices to preserve the fixed-text comparison while emphasizing sentences for which the model identifies a clear monetary policy stance. Let $\tau \in (0.5, 1)$ denote the confidence threshold. For the text model, a sentence is classified as high-confidence hawkish or dovish when that class has the highest predicted probability and the maximum class probability is at least τ. For the binary speech model, a sentence is classified as high-confidence hawkish when its predicted hawkish probability is at least τ, and as high-confidence dovish when its predicted hawkish probability is no greater than $1 - \tau$. All remaining sentences are treated as non-directional.

Let $N^{(k)}_{m,H}(\tau)$ and $N^{(k)}_{m,D}(\tau)$ denote the numbers of high-confidence hawkish and dovish sentences

in press conference m for source $k \in \{\text{Text}, \text{TTS}, \text{Hum}\}$. Conditional on observing at least one directional sentence, we define net hawkishness as

$$S_m^{(k)}(\tau) = \frac{N_{m,H}^{(k)}(\tau) - N_{m,D}^{(k)}(\tau)}{N_{m,H}^{(k)}(\tau) + N_{m,D}^{(k)}(\tau)}. \tag{3}$$

The index is positive when high-confidence hawkish sentences outnumber high-confidence dovish sentences, negative when dovish sentences predominate, and zero when the two counts are equal. Non-directional sentences do not enter the denominator. Our principal explanatory variable is the Human–TTS net hawkishness gap,

$$\Delta S_m^{\text{Hum}-\text{TTS}}(\tau) = S_m^{(\text{Hum})}(\tau) - S_m^{(\text{TTS})}(\tau). \tag{4}$$

A positive value indicates that the original full-Q&A delivery conveys a more hawkish balance of high-confidence directional signals than the TTS rendition of the same verbal content. Conversely, a negative value indicates that the actual delivery is less hawkish than the fixed-text TTS benchmark. The baseline analysis sets $\tau = 0.80$; Section 4.2 varies τ to assess the sensitivity of the results.

Our regression specification is

$$\begin{aligned} CAR_{m,a,[0,h]} = {} & \alpha_{a,h} + \beta_{1,a,h}\Delta S_m^{\text{Hum}-\text{TTS}}(\tau) + \beta_{2,a,h}S_m^{(\text{Text})}(\tau) + \gamma_{1,a,h}MP1_m + \gamma_{2,a,h}MP2_m \\ & + \eta_{1,a,h}Unemp_m + \eta_{2,a,h}PCE_m + \theta_{1,a,h}EPU_m + \theta_{2,a,h}\left(\Delta S_m^{\text{Hum}-\text{TTS}}(\tau) \times EPU_m\right) \\ & + \boldsymbol{\delta}_{a,h}^{\top}\mathbf{Z}_m + \varepsilon_{m,a,h}. \end{aligned} \tag{5}$$

Here, $MP1_m$ and $MP2_m$ measure policy-rate and expected short-rate-path surprises, respectively, from the U.S. Monetary Policy Event-Study Database (Acosta et al., 2025). $Unemp_m$ and PCE_m are the previous month's unemployment rate and year-over-year PCE inflation, respectively, and EPU_m is the daily U.S. Economic Policy Uncertainty index observed on the last available day before the meeting (Baker et al., 2016). $\mathbf{Z}_m$ contains the SEP indicator and chair-regime indicators for Bernanke and Yellen, with the Powell regime omitted. SEP equals one for press conferences accompanied by the release of the Federal Reserve's Summary of Economic Projections, and zero otherwise. All continuous macroeconomic variables are demeaned.

$\beta_{1,a,h}$ captures the association between the Human–TTS gap and market reactions when EPU is at its sample mean, while $\theta_{2,a,h}$ tests whether this association varies with policy uncertainty. All models are estimated by ordinary least squares with standard errors clustered by year. The main tables set $\tau = 0.80$, use the $[0, 1]$ event window, and report results for all ten market series. Robustness tests re-estimate Equation (5) over confidence thresholds from 0.50 to 0.95 in increments of 0.05 and event windows from $[0, 0]$ through $[0, 15]$.

4.1. Results

Tables 9 and 10 report the full specification for the baseline confidence threshold $\tau = 0.80$ and event window $[0, 1]$. Figures 3– 6 hold the threshold fixed and trace the Human–TTS coefficient and its EPU interaction across windows $[0, h]$, for $h = 0, \ldots, 15$. Figures 7 and 8 then vary both the confidence threshold and the short event window.

Treasury markets. Table 9 shows a positive, maturity-ordered relation between the voice gap and Treasury ETF returns. The coefficients for SHY (1–3Y), IEI (3–7Y), IEF (7–10Y), and TLT (20Y+) are 0.005, 0.012, 0.020, and 0.040, with year-clustered t-statistics of 1.915, 2.052, 2.301, and 3.008, respectively. Because the voice gap is a bounded index with a sample standard deviation of 0.158 (Table 7), we scale the estimates by that standard deviation rather than by a unit change, which lies well outside the observed range of $[-0.329, 0.342]$. A one-standard-deviation increase in the gap is associated with approximately 8, 19, 32, and 63 basis points of additional two-day CAR for SHY, IEI, IEF, and TLT, respectively. The coefficient rises monotonically with maturity, consistent with the greater interest-rate exposure of longer-duration securities. Expressed relative to each asset's own CAR standard deviation, however, the four effects are essentially identical, at 0.44, 0.39, 0.40, and 0.40 standard deviations. The maturity ordering therefore reflects duration scaling rather than four independent findings: the four Treasury CARs are highly correlated with one another (Table 8 reports pairwise correlations between 0.547 and 0.966, and above 0.870 for adjacent maturities), so the Treasury evidence should be read as one result observed across the curve. Textual hawkishness is insignificant in all four regressions. We do not read this as evidence that verbal content is uninformative, since the text index is uniformly dovish in this sample, with a mean of -0.805 and a maximum of -0.455 (Table 7), and therefore carries limited usable variation as a control.

The sign of the voice-gap coefficient contrasts with that of the conventional path surprise. MP2 is negative for all four Treasury ETFs and significant at the 1% level for SHY, IEI, and IEF and at the 10% level for TLT. A more-restrictive-than-expected rate path therefore lowers bond-price CARs, as standard duration logic predicts. In contrast, a larger Human–TTS gap is associated with higher bond-price CARs after controlling for MP2. This contrast weighs against treating the voice gap as another measure of the same policy-path shock. The raw correlations point the same way: in Table 8, the voice gap correlates only -0.085 with MP1 and -0.154 with MP2, so the two measures are close to orthogonal before any conditioning.

The positive sign is itself the most interesting feature of the Treasury results and warrants comment. Original full-Q&A speech assessed as more hawkish relative to the same-text TTS benchmark predicts *higher* bond prices, that is, lower yields, which is the opposite of the sign on the conventional path surprise in the same regression. A pure expected-rate-path interpretation cannot generate this pattern. One possible interpretation is a credibility or inflation-expectations channel: a Q&A interaction perceived as clearer or more resolute may coincide with greater confidence that the Committee will contain inflation, lowering expected inflation and the associated risk premium even if the expected policy path is unchanged or firmer. This mechanism carries a sharp testable implication that we do not attempt to evaluate here, namely that the effect should load on the inflation-compensation component of nominal yields rather than on real rates. It should therefore be visible in the spread between nominal Treasury and TIPS returns. Evaluating that decomposition requires a breakeven-based design and is left to future work. Alternative readings involving information quality, the composition of questions and answers, policy resolve, or term premia remain possible, and the regressions do not identify among them. The estimates should be read as conditional associations rather than causal effects of any participant's vocal delivery.

Figure 3 locates the Treasury response in the first few trading days. The positive estimates are most precise over windows ending between one and three days after the press conference; at longer horizons the confidence intervals generally widen and include zero. Neither Table 9 nor Figure 5 provides systematic evidence that EPU moderates this Treasury relation. The four $[0, 1]$ interaction estimates are insignificant, and the longer-window figure contains only isolated significant negative estimates. The Treasury evidence therefore concerns the average conditional voice-gap coefficient, not a robust uncertainty interaction.

Other asset classes. The average voice-gap relation is much less pervasive outside nominal Treasuries. In Table 10, the main effects for SPY, TIP, HYG, VIX, and VVIX are statistically indistinguishable from zero. LQD is the only exception: its coefficient is 0.034 ($t = 1.868$), significant at the 10% level and directionally consistent with the Treasury estimates. The window profiles in Figure 4 reinforce this distinction. SPY, TIP, VIX, and VVIX change sign across horizons, HYG remains positive but imprecisely estimated, and LQD exhibits only isolated 5%-level significance. By contrast, MP2 is negative and significant at the 1% level for SPY, TIP, HYG, and LQD, again separating the conventional path surprise from the voice-gap measure.

The sharper evidence for equities and volatility is state dependent. Table 10 reports a positive $\big(\text{Human–TTS}\big) \times EPU$ interaction for SPY (0.0003; $t = 2.730$) and negative interactions for VIX (-0.058; $t = -3.135$) and VVIX (-0.163; $t = -4.173$); all three are significant at the 1% level. Because EPU is measured in index points, these coefficients are small in absolute terms and are more readily interpreted when scaled. Evaluating the interaction at a one-standard-deviation increase in both the voice gap (0.158) and EPU (119.9) implies an incremental effect of approximately 57 basis points on the SPY two-day CAR, a decline of roughly 1.1 points in VIX, and a decline of roughly 3.1 points in VVIX. Relative to the sample standard deviations of the corresponding dependent variables in Table 7, each of these amounts to about 0.3 standard deviations, so the three markets move by comparable magnitudes once the uncertainty state is taken into account. The corresponding main effects are insignificant, so these estimates do not imply an unconditional equity or volatility response. Instead, as pre-conference policy uncertainty rises, the conditional association between the voice gap and SPY returns becomes more positive. In contrast, its association with changes in VIX and VVIX becomes more negative. Interactions for TIP, HYG, and LQD are insignificant.

Figure 6 shows the same state dependence across event windows. The SPY interaction remains positive throughout the plotted horizon and is significant in several short and intermediate windows. The VIX and VVIX interactions are negative and significant across many windows through roughly day nine, before attenuating at longer horizons. The combination of a more positive equity-return slope and more negative volatility-index slopes is consistent with a reduction in uncertainty or risk premia when policy uncertainty is elevated. It is less consistent with a pure tightening shock, which would ordinarily depress equity prices. This interpretation is suggestive: the interaction regressions do not separately identify clarity, interaction, information, participant-composition, or

risk-premium channels.

4.2. Sensitivity to Thresholds and Event Windows

Figure 7 shows that the nominal-Treasury result is not unique to $\tau = 0.80$ and $[0, 1]$. SHY, IEI, and IEF exhibit contiguous regions of positive significance across adjacent thresholds, concentrated in $[0, 2]$ and $[0, 3]$; TLT shows a similar positive region over short windows, although across a narrower set of thresholds. LQD provides more dispersed positive evidence, whereas the other assets do not display a comparable main-effect pattern. Precision generally weakens as the window extends to $[0, 4]$ and $[0, 5]$, indicating a short-horizon market response rather than persistent return predictability.

Appendix Tables B.3 and B.4 re-estimate the full specification over the $[0, 2]$ window and confirm that the main conclusions do not depend on the two-day horizon, with one qualification. For the Treasury ETFs, the voice-gap coefficients rise to 0.008, 0.018, 0.027, and 0.058 for SHY, IEI, IEF, and TLT. The short and intermediate maturities strengthen appreciably: SHY, IEI, and IEF are now significant at the 1% level, with t-statistics of 3.439, 3.098, and 3.097, compared with 1.915, 2.052, and 2.301 at $[0, 1]$. TLT moves in the opposite direction, weakening from the 1% to the 10% level ($t = 1.923$). The monotonic ordering of the coefficients across maturities is therefore preserved at $[0, 2]$, but the ordering of statistical precision is not, and we do not read the longest-maturity result as the most reliable of the four. The state-dependent equity and volatility findings replicate closely: the $(\text{Human–TTS}) \times EPU$ interaction remains positive for SPY and negative for VIX and VVIX, significant at the 5%, 1%, and 1% levels respectively.

Figure 8 yields an equally clear cross-specification contrast. The SPY interaction is positive across most thresholds and short event windows, while the VIX and VVIX interactions are negative across nearly all settings. Treasury interactions are largely insignificant. We therefore place weight on two coherent patterns rather than on isolated significant cells: a positive average association between the voice gap and nominal Treasury ETF returns, increasing with maturity, and an uncertainty-dependent association with equity and volatility markets. Together, the results indicate that the fixed-text voice contrast contains market-relevant information beyond textual stance and standard monetary policy surprises, while remaining associational evidence about the role of vocal delivery.

5. Conclusion

Using 89 FOMC press conferences from 2011 to 2025, this paper examines whether the vocal delivery of the full press-conference Q&A contains information beyond textual stance. The methodological contribution is to transfer monetary-policy stance labels from text to speech and compare each original sentence with a TTS rendition of identical content. Because the baseline Q&A sample includes the chair's answers, reporters' questions, and moderator remarks, the Human–TTS gap is interpreted as a fixed-text measure of the interactive Q&A's vocal delivery and composition, not as a chair-only voice effect.

We find that a larger full-Q&A Human–TTS hawkishness gap is associated with higher $[0, 1]$ CARs for U.S. Treasury ETFs across maturities, with weaker evidence for the investment-grade corporate-bond ETF. These associations remain distinct from conventional monetary-policy surprises, particularly the expected-path surprise, which enters negatively across most fixed-income specifications. The voice gap therefore contains market-relevant information not captured by textual stance or standard policy-rate and path surprises.

Equity and volatility responses are state-dependent. When economic policy uncertainty is high, a larger gap is associated with higher SPY CARs and lower VIX and VVIX responses. The acoustic analysis further shows that WavLM-implied Q&A hawkishness is systematically related to vocal perturbation, pitch, and loudness control. These patterns characterize the interactive communication environment and may reflect both delivery and speaker composition; they should not be interpreted as causal effects of the chair's individual voice.

Overall, the evidence indicates that FOMC press conferences are multimodal communication events in which both verbal content and the vocal structure of the Q&A matter. The fixed-text Human–TTS contrast is most informative for Treasury-market CARs and for equity and volatility responses during periods of elevated policy uncertainty.

Table 1: Data-Processing Pipeline and Sample Size at Each Stage

Stage	Count	Description or filter
1. Raw source	89 conferences	FOMC press conferences from April 27, 2011, to December 10, 2025.
2. ASR and sentence segmentation	42,819 sentences	Whisper large-v3 transcription followed by sentence assembly.
3. Cleaning	35,619 sentences	Remove audio clips lasting at most 1.0 second and sentences containing at most three words.
4. High-confidence binary sample	8,256 sentences	Hawkish or dovish text labels with classification confidence of at least 0.80.
5. Final training set	8,687 windows	High-confidence sentences after dividing clips longer than 20 seconds into overlapping 20-second windows with a two-second overlap.
6. Full-sample inference inputs	36,695 windows	All 35,619 cleaned sentences after applying the same slicing rule.
7. Sentence-level inference	35,619 sentences	Average window-level hawkish probabilities within each original sentence to recover one prediction per cleaned sentence.

Table 2: WavLM Cross-Validation and Classification Performance

Panel A. Stratified group five-fold cross-validation				
Fold	AUC	Accuracy	Macro F1	
Fold 1	0.9209	0.8495	0.8491	
Fold 2	0.9076	0.8297	0.8293	
Fold 3	0.9065	0.8363	0.8349	
Fold 4	0.9065	0.8317	0.8283	
Fold 5	0.9174	0.8350	0.8349	
Pooled out-of-fold	0.9125	0.8362	0.8362	
Panel B. Class-specific out-of-fold performance				
Class	Precision	Recall	F1	Audio windows
Dovish	0.8336	0.8350	0.8343	4,291
Hawkish	0.8387	0.8374	0.8380	4,396
Macro average	0.8362	0.8362	0.8362	8,687
Weighted average	0.8362	0.8362	0.8362	8,687

Notes: Panel A reports predictive performance from stratified group five-fold cross-validation, with press-conference date defining the group. All sentences and audio windows from the same conference are assigned to the same fold. The pooled row is calculated from the combined held-out predictions across the five folds. Panel B reports class-specific performance for the 8,687 audio windows. Macro averages weight the hawkish and dovish classes equally, whereas weighted averages use their respective numbers of audio windows. AUC denotes the area under the receiver operating characteristic curve.

Table 3: Paired t-Tests of Human vs. TTS Hawkishness (Statement and Q&A)

Level	Section	N	Human mean	TTS mean	Mean gap	t-stat
Sentence	Statement	4,948	0.5597	0.5378	0.0218	10.2073***
Meeting	Statement	89	0.5714	0.5556	0.0158	2.6805***
Sentence	Q&A	30,671	0.4225	0.4125	0.0100	9.9376***
Meeting	Q&A	89	0.4589	0.4469	0.0120	2.2429**

Notes: The table reports paired tests. The mean gap is defined as human minus TTS hawkishness. At the sentence level, each original human-audio sentence is paired with the TTS rendition of the identical text. The Q&A human sample pools chair answers, reporter questions, and moderator remarks. At the meeting level, sentence-level probabilities are first aggregated using audio-duration weights. ***, **, and * denote $p < 0.01$, $p < 0.05$, and $p < 0.10$, respectively.

Table 4: Paired t-Tests of Acoustic-Feature Gaps (Human Minus TTS)

Feature	Sentence: Statement	Sentence: Q&A	Meeting: Statement	Meeting: Q&A
Mean F0	0.7352***	0.3448***	0.8181	0.7217
F0 STD	-1.5179***	-1.1863***	-1.2681***	-0.8150
F0 Range	25.4286***	35.1062***	30.0508***	45.0890***
Mean Loudness	-7.6562***	-7.1236***	-8.1323***	-8.0002***
Loudness STD	-5.8676***	-7.9843***	-5.1849***	-6.3064***
Jitter	-0.0017***	-0.0021***	-0.0019***	-0.0025***
Shimmer	0.0172***	0.0217***	0.0167***	0.0182***
HNR	-1.6088***	-1.6422***	-1.5545***	-1.1801***

Notes: Entries report mean acoustic-feature gaps, defined as human speech minus the same-text TTS audio. The Q&A human sample pools chair answers, reporter questions, and moderator remarks. Sentence-level gaps compare paired human and TTS renditions of identical text; meeting-level gaps are constructed from audio-duration-weighted sentence-level features. Significance is based on paired t-tests. Mean F0 measures the overall pitch level; F0 STD measures within-clip pitch variation; F0 Range is the difference between the maximum and minimum F0; Mean Loudness measures average speech intensity; Loudness STD measures within-clip intensity variation; Jitter measures cycle-to-cycle irregularity in fundamental frequency; Shimmer measures cycle-to-cycle irregularity in amplitude; and HNR is the harmonics-to-noise ratio, with higher values indicating a clearer, more periodic, and more stable voice. ***, **, and * denote $p < 0.01$, $p < 0.05$, and $p < 0.10$, respectively.

Table 5: Market Assets and Response Construction

Ticker	Instrument and type	Market dimension
SHY	iShares 1–3 Year Treasury Bond ETF	Short-maturity U.S. Treasury market.
IEI	iShares 3–7 Year Treasury Bond ETF	Short-to-intermediate U.S. Treasury market.
IEF	iShares 7–10 Year Treasury Bond ETF	Intermediate-to-long U.S. Treasury market.
TLT	iShares 20+ Year Treasury Bond ETF	Long-maturity U.S. Treasury market.
SPY	SPDR S&P 500 ETF	U.S. large-cap equity market.
TIP	iShares TIPS Bond ETF	U.S. inflation-protected Treasury market.
HYG	iShares iBoxx $ High Yield Corporate Bond ETF	U.S. below-investment-grade corporate-bond market.
LQD	iShares iBoxx $ Investment Grade Corporate Bond ETF	U.S. investment-grade corporate-bond market.
VIX	Cboe Volatility Index	Short-term implied volatility of the U.S. equity market.
VVIX	Cboe VVIX Index	Implied volatility of VIX options.

Notes: SHY, IEI, IEF, TLT, SPY, TIP, HYG, and LQD are exchange-traded funds. For these assets, daily returns are calculated from adjusted closing prices as $r_{t,a} = \log\left(\frac{P_{t,a}}{P_{t-1,a}}\right)$. VIX and VVIX are Cboe indices rather than traded ETFs; their daily responses are therefore measured as changes in index levels, $r_{t,a} = P_{t,a} - P_{t-1,a}$. Expected market responses are estimated separately for each asset using the 60 trading days preceding the event window. Daily abnormal responses are the differences between realized and expected responses, and cumulative abnormal responses are obtained by summing them over the indicated event window. Accordingly, CAR is interpreted as a cumulative abnormal log return for the eight ETFs and as a cumulative abnormal index-level change for VIX and VVIX. *Source:* Yahoo!Finance.

Table 6: Variables Used in the CAR Regressions

Variable	Definition
$CAR_{m,a,[0,h]}$	Cumulative abnormal response of asset a over event window $[0, h]$ following press conference m: cumulative abnormal log return for the eight ETFs and cumulative abnormal index-level change for VIX and VVIX.
$\Delta S_m^{\text{Hum}-\text{TTS}}(\tau)$	Full-Q&A Human–TTS voice gap: high-confidence net hawkishness of all original Q&A speech including chair answers, reporter questions, and moderator remarks, minus that of the same-text TTS audio.
$S_m^{(\text{Text})}(\tau)$	High-confidence frequency-based net hawkishness constructed from FOMC-RoBERTa text predictions.
$MP1_m$	High-frequency surprise in the current policy rate at press conference m, constructed from federal funds futures.
$MP2_m$	High-frequency revision in the expected policy rate at the next scheduled FOMC meeting, constructed from federal funds futures.
$Unemp_m$	Unemployment rate.
PCE_m	Year-over-year PCE inflation, $100(PCEPI_t/PCEPI_{t-12} - 1)$.
EPU_m	U.S. Economic Policy Uncertainty index.
$\Delta S_m^{\text{Hum}-\text{TTS}}(\tau) \times EPU_m$	Interaction between the Human–TTS voice gap and economic policy uncertainty.
SEP_m	Indicator equal to one for meetings with a Summary of Economic Projections release.
$Bernanke_m$	Indicator equal to one for press conferences held during the Bernanke tenure.
$Yellen_m$	Indicator equal to one for press conferences held during the Yellen tenure; the Powell tenure is the omitted category.

Notes: The construction of $S_m^{(k)}(\tau)$ is described in the text; the baseline sets $\tau = 0.80$. Market series used to construct CARs are reported in Table 5. $MP1_m$ and $MP2_m$ are from the Press Conferences file of the U.S. Monetary Policy Event-Study Database (USMPD), Federal Reserve Bank of San Francisco. UNRATE, PCEPI, and USEPUINDXD are from FRED, and SEP_m is constructed from Federal Reserve FOMC meeting materials. Let t_m denote the press-conference date. $MP1_m$ and $MP2_m$ are event-specific measures matched directly to t_m, without averaging or demeaning. $Unemp_m$ and PCE_m use the preceding calendar month, whereas EPU_m is the latest available daily observation strictly before t_m. These three controls are mean-centered across the 89 conferences, $X_m = X_m^{\text{raw}} - \bar{X}$, rather than averaged between FOMC meetings. For CAR construction, the normal market response is the arithmetic mean over the 60 available trading days immediately preceding t_m; this window may span earlier FOMC meetings.

Table 7: Descriptive Statistics of Regression Variables

Variable	Mean	SD	Min.	P25	Median	P75	Max.
Panel A. Dependent variables							
$CAR_{m,SHY,[0,1]}$ [1–3Y]	0.0006	0.0018	-0.0024	-0.0005	0.0001	0.0013	0.0072
$CAR_{m,IEI,[0,1]}$ [3–7Y]	0.0012	0.0048	-0.0095	-0.0021	0.0013	0.0034	0.0148
$CAR_{m,IEF,[0,1]}$ [7–10Y]	0.0015	0.0078	-0.0179	-0.0034	0.0016	0.0056	0.0238
$CAR_{m,TLT,[0,1]}$ [20Y+]	0.0022	0.0157	-0.0308	-0.0075	0.0028	0.0095	0.0483
$CAR_{m,SPY,[0,1]}$ [Equity]	-0.0008	0.0192	-0.0710	-0.0094	-0.0001	0.0098	0.0441
$CAR_{m,TIP,[0,1]}$ [TIPS]	0.0009	0.0088	-0.0347	-0.0053	0.0018	0.0058	0.0242
$CAR_{m,HYG,[0,1]}$ [HY]	0.0009	0.0105	-0.0445	-0.0027	0.0005	0.0044	0.0300
$CAR_{m,LQD,[0,1]}$ [IG]	0.0017	0.0103	-0.0551	-0.0025	0.0014	0.0069	0.0279
$CAR_{m,VIX,[0,1]}$ [Vol.]	0.0367	3.2796	-7.0893	-1.3460	-0.3580	0.6430	16.5570
$CAR_{m,VVIX,[0,1]}$ [VoV]	-0.6658	9.8930	-27.2177	-6.3240	-1.3427	2.9383	42.6093
Panel B. Explanatory variables							
Human–TTS	0.0307	0.1578	-0.3294	-0.0853	0.0343	0.1544	0.3421
Text	-0.8048	0.1048	-0.9686	-0.8792	-0.8289	-0.7568	-0.4550
MP1	-0.0001	0.0053	-0.0332	0.0000	0.0000	0.0000	0.0125
MP2	-0.0022	0.0207	-0.0900	-0.0065	0.0000	0.0050	0.0714
Unemp	0.0000	1.9160	-1.7124	-1.3124	-0.9124	0.9876	8.0876
PCE	0.0000	1.7272	-2.3814	-1.1473	-0.3717	0.2353	4.6914
EPU	0.0000	119.9051	-132.1496	-73.6896	-32.2496	36.2604	461.3704
$(\text{Human–TTS}) \times EPU$	2.7724	18.6834	-43.2832	-5.3918	0.1576	7.3881	87.6970
SEP	0.6404	0.4826	0.0000	0.0000	1.0000	1.0000	1.0000
Bernanke	0.1348	0.3435	0.0000	0.0000	0.0000	0.0000	1.0000
Yellen	0.1798	0.3862	0.0000	0.0000	0.0000	0.0000	1.0000

Notes: Statistics are calculated at the FOMC meeting level for the Q&A sample using the main high-confidence threshold $\tau = 0.80$. The sample contains 89 press conferences, and no variable has missing observations. The ten dependent variables are $CAR_{m,a,[0,1]}$. For SPY, SHY, IEI, IEF, TLT, TIP, HYG, and LQD, CAR is a cumulative abnormal log return constructed from adjusted closing prices without subtracting the risk-free rate; for VIX and VVIX, it is a cumulative abnormal change in the index level. In Panel B, Human–TTS is the high-confidence net hawkishness of all original full-Q&A speech—chair answers, reporter questions, and moderator remarks—minus that of the same-text TTS audio; Text is the textual stance control; MP1 and MP2 are monetary-policy surprises; Unemp and PCE are macroeconomic controls; EPU is the policy-uncertainty index; and SEP, Bernanke, and Yellen are indicator variables. Bernanke and Yellen are chair-regime indicators, with the Powell regime omitted. Unemp, PCE, and EPU are mean-centered. HY and IG denote high-yield and investment-grade credit; Vol. and VoV denote equity volatility and volatility of volatility. SD is the sample standard deviation; P25 and P75 are the 25th and 75th percentiles. The regression intercept is excluded because it is fixed at one and has no meaningful dispersion.

Table 8: Pearson Correlations among Regression Variables

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)
CAR-SHY	1.000																				
CAR-IEI	0.925	1.000																			
CAR-IEF	0.827	0.966	1.000																		
CAR-TLT	0.547	0.737	0.870	1.000																	
CAR-SPY	0.174	0.131	0.070	-0.051	1.000																
CAR-TIP	0.725	0.809	0.793	0.639	0.388	1.000															
CAR-HYG	0.347	0.303	0.242	0.107	0.885	0.544	1.000														
CAR-LQD	0.612	0.677	0.682	0.631	0.581	0.844	0.750	1.000													
CAR-VIX	-0.039	-0.007	0.037	0.080	-0.827	-0.301	-0.735	-0.558	1.000												
CAR-VVIX	0.081	0.097	0.113	0.129	-0.760	-0.131	-0.599	-0.348	0.873	1.000											
Human–TTS	0.362	0.348	0.337	0.292	-0.036	0.273	0.086	0.316	-0.028	0.075	1.000										
Text	0.133	0.159	0.161	0.161	0.076	0.196	0.127	0.248	-0.165	-0.067	0.428	1.000									
MP1	-0.211	-0.175	-0.160	-0.125	-0.095	-0.258	-0.135	-0.231	0.111	-0.015	-0.085	-0.022	1.000								
MP2	-0.509	-0.422	-0.379	-0.244	-0.152	-0.350	-0.234	-0.388	0.069	-0.013	-0.154	-0.288	0.245	1.000							
Unemp	-0.231	-0.197	-0.198	-0.153	-0.105	-0.092	-0.096	-0.151	0.125	0.112	-0.537	-0.422	0.072	0.158	1.000						
PCE	0.167	0.150	0.111	0.003	0.036	0.102	0.083	0.135	-0.108	-0.068	0.456	0.689	0.005	-0.294	-0.376	1.000					
EPU	-0.149	-0.170	-0.166	-0.167	0.005	-0.170	-0.031	-0.172	0.005	-0.032	0.148	-0.176	0.050	0.182	0.052	-0.023	1.000				
$(\text{Human–TTS}) \times EPU$	-0.218	-0.223	-0.221	-0.207	0.179	-0.047	0.133	-0.034	-0.277	-0.279	-0.136	-0.143	-0.066	0.225	0.049	-0.170	0.510	1.000			
SEP	-0.055	-0.123	-0.103	-0.028	-0.161	-0.058	-0.155	-0.098	0.064	0.007	-0.239	-0.064	-0.052	-0.199	0.042	-0.222	-0.147	0.047	1.000		
Bernanke	-0.123	-0.134	-0.186	-0.219	0.043	-0.088	0.006	-0.093	-0.051	-0.022	-0.387	-0.231	0.055	0.093	0.612	-0.153	-0.047	-0.081	0.022	1.000	
Yellen	-0.125	-0.106	-0.059	0.020	0.125	-0.029	0.105	0.042	-0.197	-0.184	-0.514	-0.054	-0.057	0.007	0.020	-0.403	-0.274	0.265	0.351	-0.185	1.000

Variable key:

(1) $CAR_{m,SHY,[0,1]}$ [1–3Y]
(2) $CAR_{m,IEI,[0,1]}$ [3–7Y]
(3) $CAR_{m,IEF,[0,1]}$ [7–10Y]
(4) $CAR_{m,TLT,[0,1]}$ [20Y+]
(5) $CAR_{m,SPY,[0,1]}$ [Equity]
(6) $CAR_{m,TIP,[0,1]}$ [TIPS]
(7) $CAR_{m,HYG,[0,1]}$ [HY]
(8) $CAR_{m,LQD,[0,1]}$ [IG]
(9) $CAR_{m,VIX,[0,1]}$ [Vol.]
(10) $CAR_{m,VVIX,[0,1]}$ [VoV]
(11) Human–TTS
(12) Text
(13) MP1
(14) MP2
(15) Unemp
(16) PCE
(17) EPU
(18) $(\text{Human–TTS}) \times EPU$
(19) SEP
(20) Bernanke
(21) Yellen

Notes: The table reports lower-triangular Pearson correlation coefficients for the Q&A main-regression variables using $\tau = 0.80$ and $CAR_{m,a,[0,1]}$. Correlations use pairwise complete observations; the pairwise sample size is 89 throughout. The ten CAR variables are asset-specific dependent variables. Human–TTS is the high-confidence net hawkishness of all original full-Q&A speech—chair answers, reporter questions, and moderator remarks—minus that of the same-text TTS audio. HY and IG denote high-yield and investment-grade credit; Vol. and VoV denote equity volatility and volatility of volatility. Unemp, PCE, and EPU are mean-centered; SEP is an indicator, and Bernanke and Yellen are chair-regime indicators with the Powell regime omitted. The regression intercept is excluded because it has zero variance and therefore no defined correlation.

Table 9: $CAR[0,1]$ Regressions for Treasury ETFs

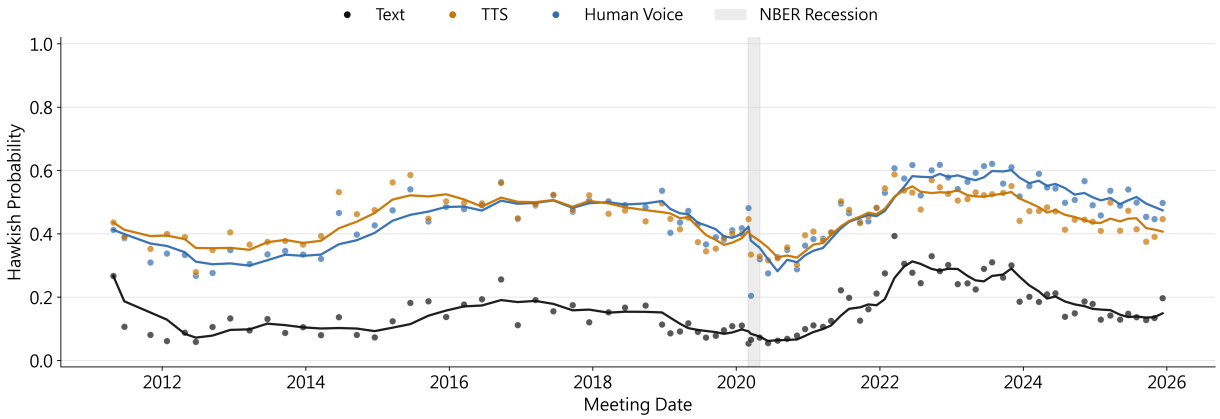
	SHY [1–3Y]	IEI [3–7Y]	IEF [7–10Y]	TLT [20Y+]
Human–TTS	0.005*	0.012**	0.020**	0.040***
	(1.915)	(2.052)	(2.301)	(3.008)
Text	-0.003	-0.005	-0.004	0.017
	(-0.856)	(-0.456)	(-0.252)	(0.678)
MP1	-0.019	-0.041	-0.060	-0.126
	(-0.737)	(-0.568)	(-0.496)	(-0.460)
MP2	-0.043***	-0.095***	-0.135***	-0.150*
	(-10.835)	(-8.946)	(-5.670)	(-1.702)
Unemp	-0.000	0.000	0.000	0.001
	(-1.304)	(0.074)	(0.382)	(0.782)
PCE	-0.000	-0.000	-0.001	-0.003**
	(-0.185)	(-0.449)	(-0.808)	(-2.103)
EPU	-0.000	-0.000	-0.000	-0.000
	(-0.855)	(-0.950)	(-0.778)	(-0.551)
$(\text{Human–TTS}) \times EPU$	-0.000	-0.000	-0.000	-0.000
	(-0.380)	(-0.483)	(-0.884)	(-1.420)
SEP	-0.001	-0.002*	-0.003	-0.003
	(-1.110)	(-1.654)	(-1.387)	(-0.572)
Bernanke	0.001	0.001	-0.001	-0.006
	(1.193)	(0.329)	(-0.295)	(-1.226)
Yellen	0.001	0.001	0.003	0.006
	(0.572)	(0.586)	(0.738)	(0.910)
Constant	-0.002	-0.002	-0.001	0.016
	(-0.649)	(-0.254)	(-0.069)	(0.787)
Observations	89	89	89	89
R^2	0.405	0.335	0.303	0.246
Adjusted R^2	0.320	0.240	0.204	0.138

Notes: The dependent variable is the cumulative abnormal log return over the event window $[0,1]$, denoted $CAR_{m,a,[0,1]}$. SHY, IEI, IEF, and TLT track Treasury securities with maturities of 1–3, 3–7, 7–10, and more than 20 years, respectively. Human–TTS is the high-confidence net hawkishness of all original full-Q&A speech—chair answers, reporter questions, and moderator remarks—minus that of the same-text TTS audio. Text is the corresponding high-confidence net hawkishness constructed from verbal content. MP1 and MP2 measure the current policy-rate surprise and the revision in the expected policy rate at the next scheduled FOMC meeting. Unemp and PCE are mean-centered unemployment and year-over-year PCE inflation; EPU is the mean-centered Economic Policy Uncertainty index. SEP indicates the release of a Summary of Economic Projections. Bernanke and Yellen are chair-regime indicators, with the Powell regime omitted. Year-clustered t-statistics are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

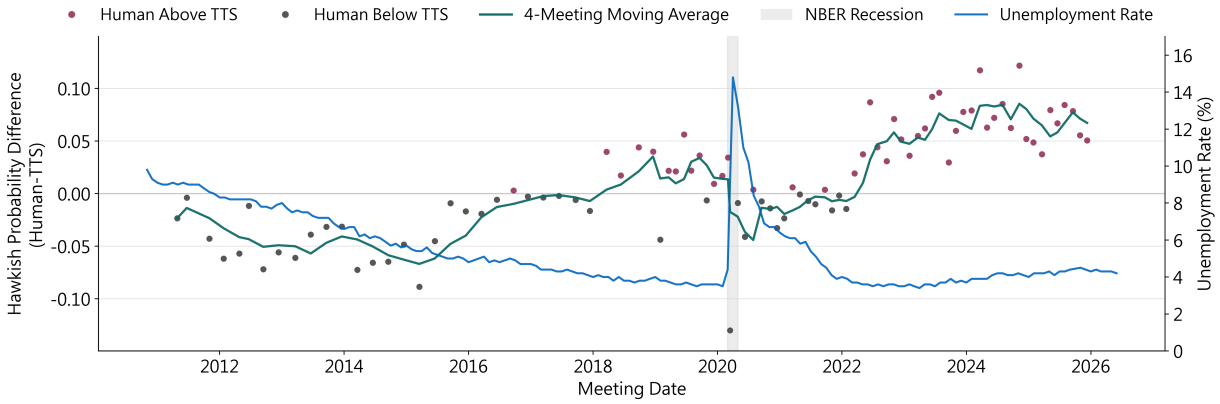
Table 10: $CAR[0, 1]$ Regressions for Equity, Inflation, Credit, and Volatility Markets

	SPY [Equity]	TIP [TIPS]	HYG [HY]	LQD [IG]	VIX [Vol.]	VVIX [VoV]
Human–TTS	-0.007	0.021	0.013	0.034*	-4.057	1.116
	(-0.269)	(1.443)	(0.803)	(1.868)	(-0.798)	(0.109)
Text	0.006	0.008	-0.001	0.004	-0.144	5.744
	(0.205)	(0.500)	(-0.066)	(0.278)	(-0.028)	(0.394)
MP1	-0.072	-0.225**	-0.061	-0.160	26.418	-99.775
	(-0.251)	(-2.075)	(-0.370)	(-1.033)	(0.482)	(-0.640)
MP2	-0.238***	-0.138***	-0.155***	-0.185***	13.218	13.170
	(-2.912)	(-4.436)	(-4.398)	(-4.075)	(0.715)	(0.240)
Unemp	-0.003***	0.001	-0.001**	0.001	0.321***	1.244***
	(-5.789)	(0.766)	(-2.471)	(0.541)	(2.578)	(3.798)
PCE	-0.001	-0.001	-0.000	-0.001	-0.283	-1.049
	(-0.246)	(-0.950)	(-0.039)	(-0.855)	(-0.665)	(-1.060)
EPU	-0.000	-0.000	-0.000	-0.000	0.002	0.004
	(-0.428)	(-1.381)	(-0.623)	(-1.365)	(0.889)	(0.825)
(Human–TTS) $\times$ EPU	0.000***	0.000	0.000	0.000	-0.058***	-0.163***
	(2.730)	(1.117)	(1.617)	(0.988)	(-3.135)	(-4.173)
SEP	-0.012**	-0.002	-0.006**	-0.005***	1.097	1.536
	(-2.250)	(-1.420)	(-2.570)	(-2.644)	(1.005)	(0.612)
Bernanke	0.014**	0.000	0.007	0.004	-3.551***	-7.004***
	(2.259)	(0.036)	(1.555)	(1.251)	(-3.730)	(-3.060)
Yellen	0.008	0.001	0.007	0.007	-3.200**	-5.877*
	(1.005)	(0.122)	(1.392)	(1.502)	(-2.027)	(-1.674)
Constant	0.007	0.007	0.001	0.004	0.588	5.412
	(0.298)	(0.556)	(0.065)	(0.351)	(0.133)	(0.427)
Observations	89	89	89	89	89	89
R^2	0.195	0.280	0.206	0.362	0.252	0.170
Adjusted R^2	0.080	0.177	0.093	0.270	0.145	0.051

Notes: The dependent variable is $CAR_{m,a,[0,1]}$. For SPY, TIP, HYG, and LQD, CAR is a cumulative abnormal log return; for VIX and VVIX, it is a cumulative abnormal change in the index level. Human–TTS is the high-confidence net hawkishness of all original full-Q&A speech—chair answers, reporter questions, and moderator remarks—minus that of the same-text TTS audio. Text is high-confidence net textual hawkishness. MP1 and MP2 are the current-rate and expected-path surprises. Unemp, PCE, and EPU are mean-centered. SEP is the Summary of Economic Projections indicator. Bernanke and Yellen are chair-regime indicators, with the Powell regime omitted. Year-clustered t-statistics are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels.



(a) Meeting-Level Text, TTS, and Human Hawkishness over Time



(b) Meeting-Level Human–TTS Hawkishness Gap over Time

Figure 1: Meeting-Level Hawkishness and Human–TTS Gaps over Time

Notes: This figure reports the time-series variation in meeting-level hawkishness during the Q&A. Panel A plots the hawkishness implied by the sentence text, the TTS rendition of the same text, and the original full-Q&A speech. Sentence-level probabilities are aggregated to the meeting level using audio-duration weights. Panel B plots the corresponding Human–TTS gap, defined as human-speech hawkishness minus TTS hawkishness. Positive values indicate that the original human delivery is assessed as more hawkish than the standardized TTS rendition of identical verbal content; negative values indicate the opposite. The TTS benchmark holds the sentence text fixed, so variation in the gap is interpreted as a delivery-related contrast rather than a difference in verbal content.

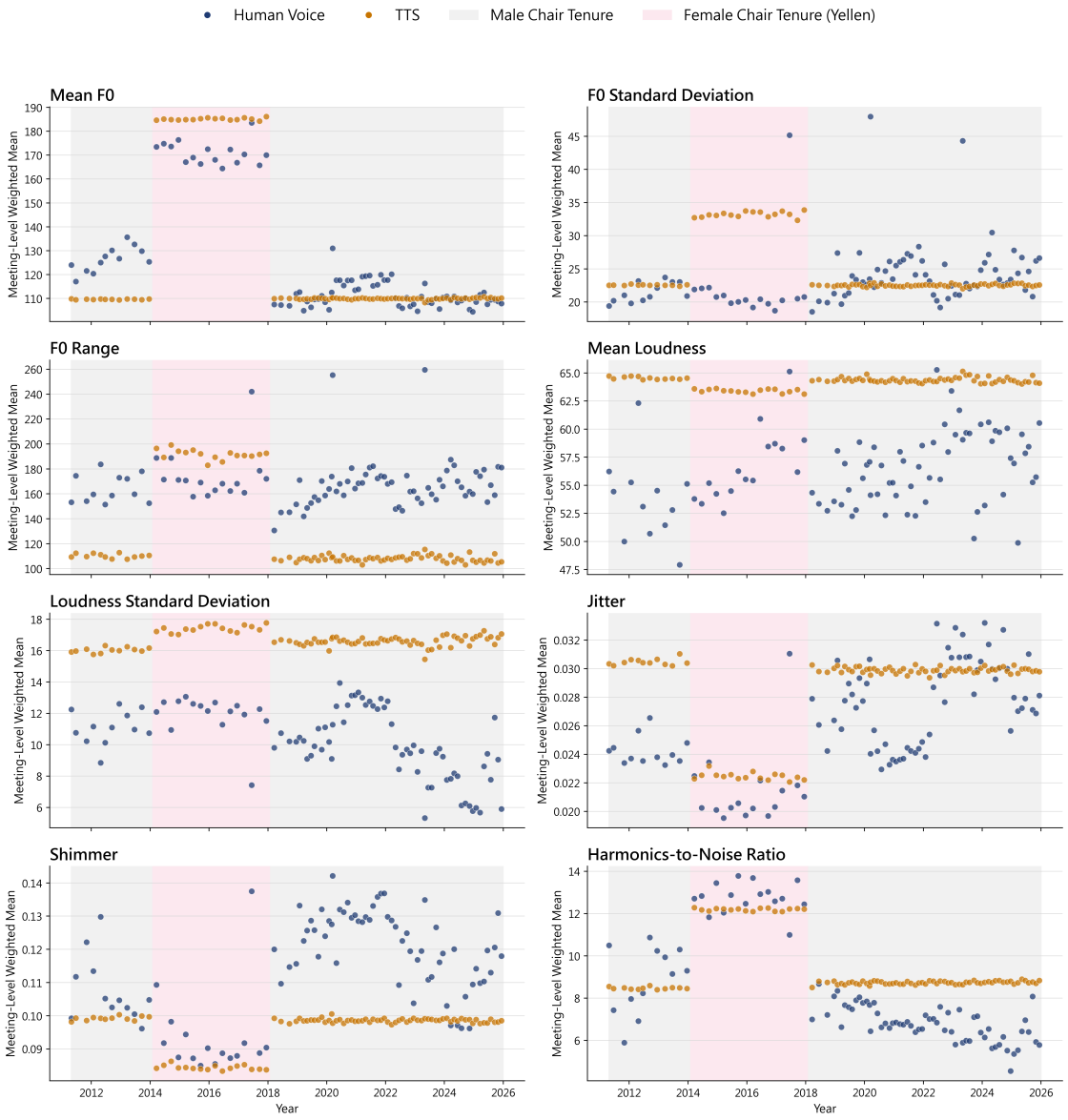
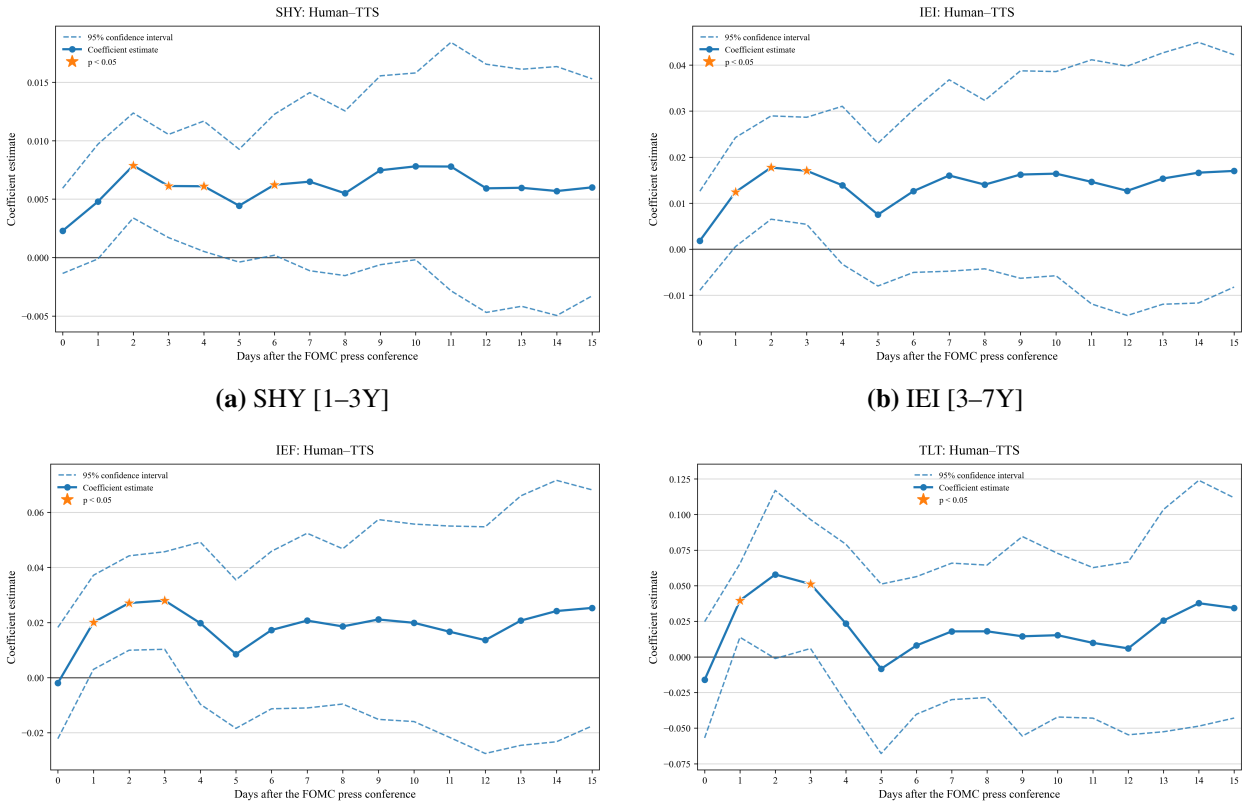


Figure 2: Meeting-Level Acoustic Features of Original and TTS Speech during the Full Q&A

Notes: This figure plots the meeting-level acoustic features of the original full-Q&A speech and the corresponding text-to-speech (TTS) renditions. The original speech pools chair answers, reporter questions, and moderator remarks. Human and TTS audio are constructed from identical sentence text, and sentence-level acoustic features are aggregated to the meeting level using audio-duration weights. The features include mean pitch, pitch variation, pitch range, mean loudness, loudness variation, jitter, shimmer, and the harmonics-to-noise ratio. Differences between the two series therefore reflect delivery and speaker composition relative to the standardized TTS benchmark, holding verbal content fixed.



(a) SHY [1–3Y]

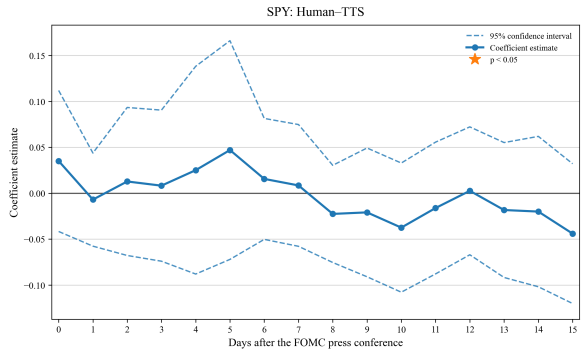
(b) IEI [3–7Y]

(c) IEF [7–10Y]

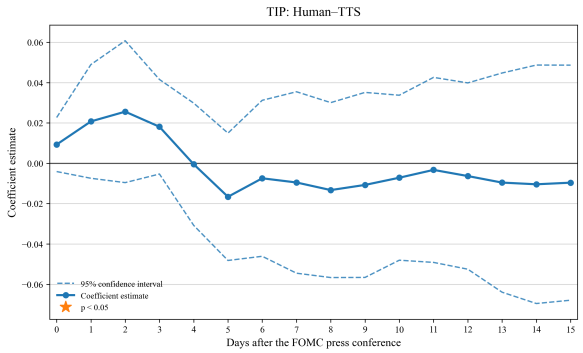
(d) TLT [20Y+]

Figure 3: Human–TTS Coefficient Estimates across U.S. Treasury ETF Maturities

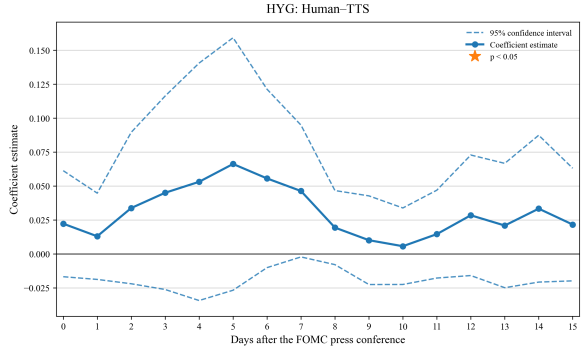
Notes: The figure plots coefficient estimates for the meeting-level Human–TTS net hawkishness gap across event windows for four U.S. Treasury ETFs. SHY, IEI, IEF, and TLT track Treasury securities with maturities of 1–3, 3–7, 7–10, and more than 20 years, respectively. The dependent variable is the cumulative abnormal log return over the indicated post-conference event window. Solid lines report the coefficient estimates, and dashed lines report the corresponding 95% confidence intervals. Positive coefficients indicate that a larger Human–TTS hawkishness gap is associated with higher Treasury ETF cumulative abnormal returns. All specifications control for text stance, monetary policy surprises, macroeconomic conditions, EPU, the $(\text{Human–TTS}) \times EPU$ interaction, the SEP indicator, and chair-regime indicators. Orange stars mark estimates with $p < 0.05$.



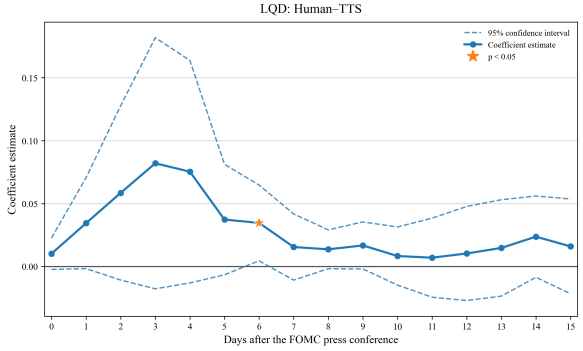
(a) SPY [Equity]



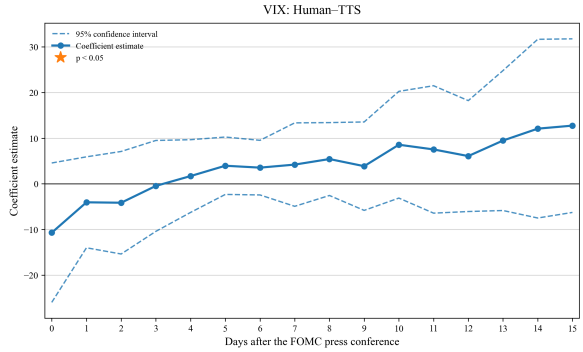
(b) TIP [TIPS]



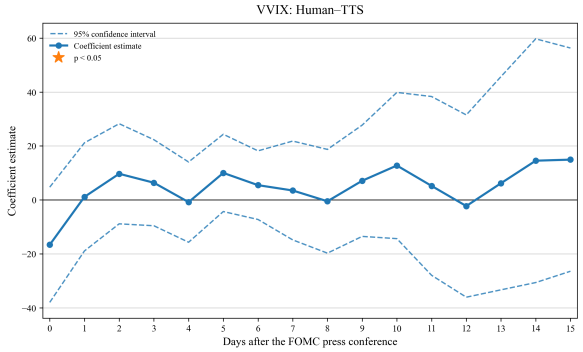
(c) HYG [HY]



(d) LQD [IG]



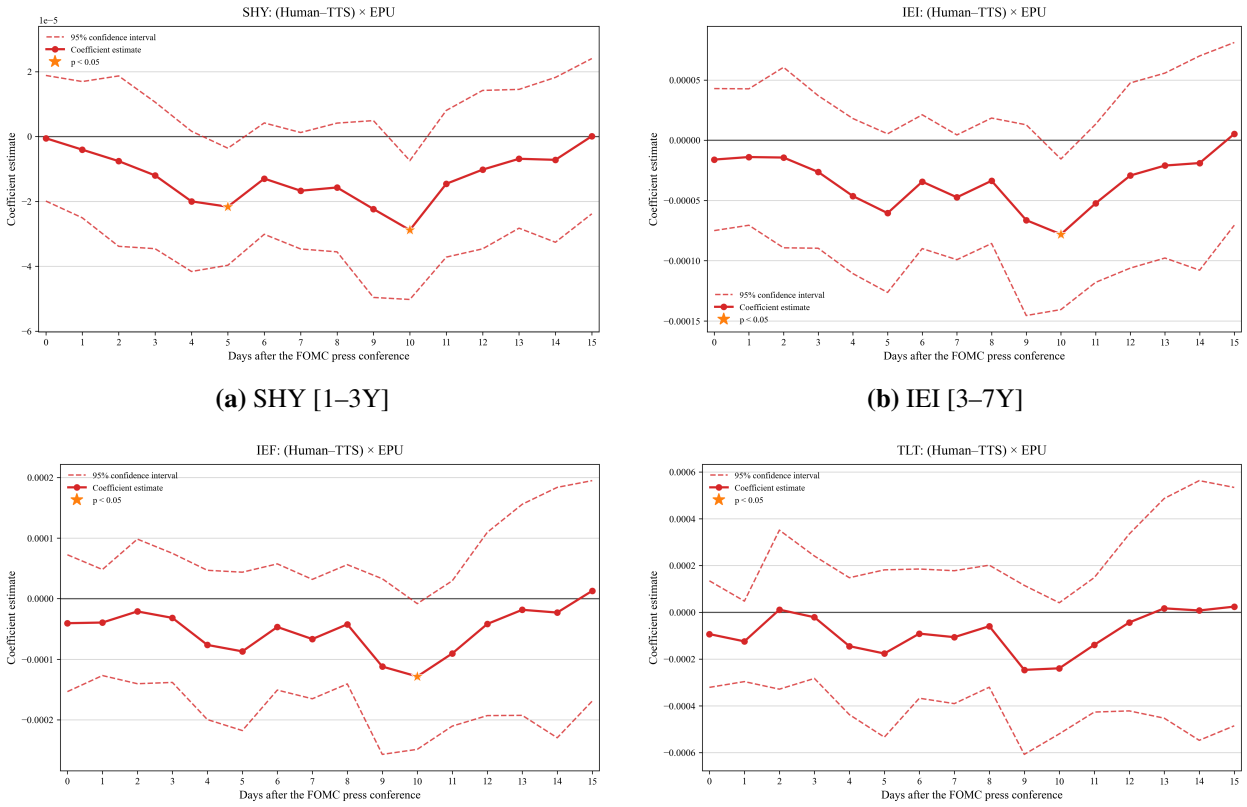
(e) VIX [Vol.]



(f) VVIX [VoV]

Figure 4: Human–TTS Coefficient Estimates across Equity, Inflation, Credit, and Volatility Markets

Notes: The figure plots coefficient estimates for the meeting-level Human–TTS net hawkishness gap from Specification (5) across post-conference event windows. SPY represents the U.S. large-cap equity market; TIP represents inflation-protected Treasury securities; HYG and LQD represent high-yield and investment-grade corporate credit; VIX measures expected equity-market volatility; and VVIX measures the implied volatility of VIX options. For SPY, TIP, HYG, and LQD, the dependent variable is the cumulative abnormal log return over the indicated event window. For VIX and VVIX, it is the cumulative abnormal change in the index level. Solid lines report coefficient estimates, and dashed lines report the corresponding 95% confidence intervals. All specifications control for text stance, monetary policy surprises, macroeconomic conditions, EPU, the $(\text{Human–TTS}) \times EPU$ interaction, the SEP indicator, and chair-regime indicators. HY and IG denote high-yield and investment-grade credit; Vol. and VoV denote equity volatility and volatility of volatility. Orange stars mark estimates with $p < 0.05$.



(a) SHY [1–3Y]

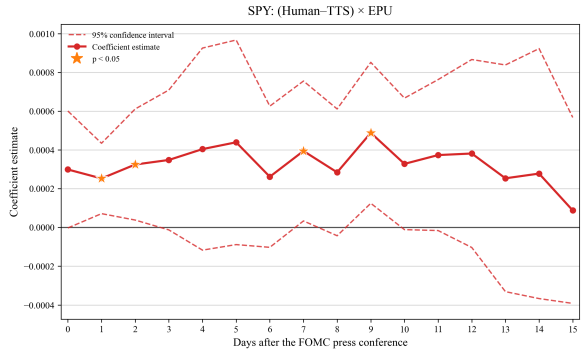
(b) IEI [3–7Y]

(c) IEF [7–10Y]

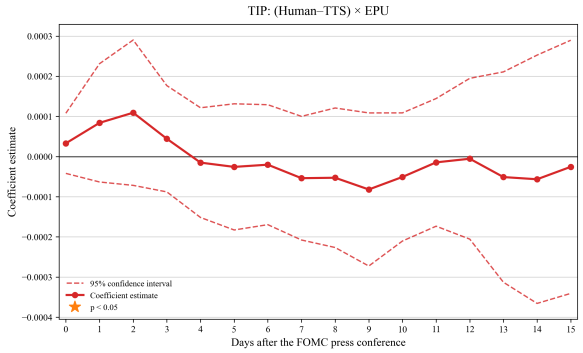
(d) TLT [20Y+]

Figure 5: Interaction-Term Coefficient Estimates across U.S. Treasury ETF Maturities

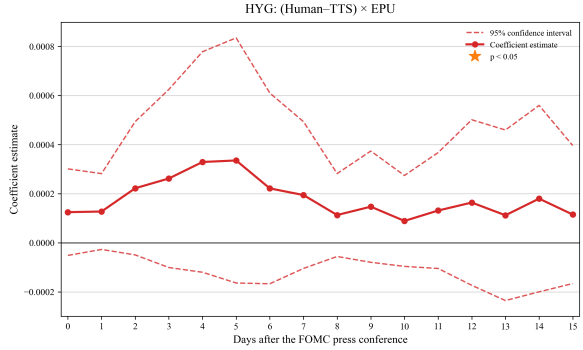
Notes: The figure plots coefficient estimates for the interaction term $(\text{Human–TTS}) \times EPU$ from Specification (5) across post-conference event windows for four U.S. Treasury ETFs. SHY, IEI, IEF, and TLT track Treasury securities with maturities of 1–3, 3–7, 7–10, and more than 20 years, respectively. The dependent variable is the cumulative abnormal log return over the indicated event window. Solid lines report the coefficient estimates, and dashed lines report the corresponding 95% confidence intervals. A positive coefficient indicates that the association between the Human–TTS hawkishness gap and Treasury ETF returns becomes more positive when economic policy uncertainty is higher. All specifications control for text stance, monetary policy surprises, macroeconomic conditions, the Human–TTS main effect, EPU, the SEP indicator, and chair-regime indicators. Orange stars mark estimates with $p < 0.05$.



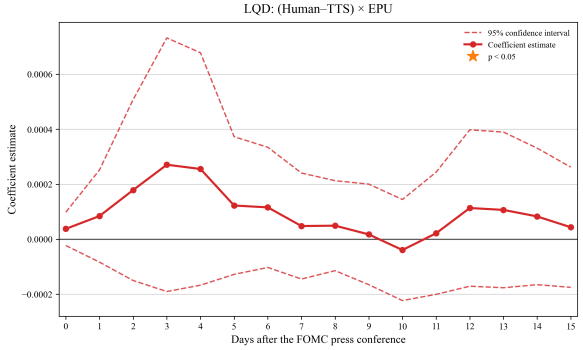
(a) SPY [Equity]



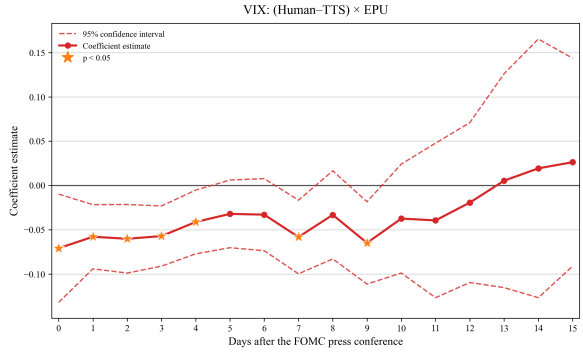
(b) TIP [TIPS]



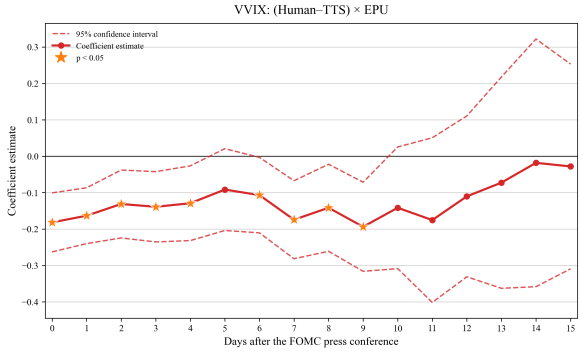
(c) HYG [HY]



(d) LQD [IG]



(e) VIX [Vol.]



(f) VVIX [VoV]

Figure 6: Interaction-Term Coefficient Estimates across Equity, Inflation, Credit, and Volatility Markets

Notes: The figure plots coefficient estimates for the interaction term $(\text{Human–TTS}) \times EPU$ from Specification (5) across post-conference event windows. SPY represents the U.S. large-cap equity market; TIP represents inflation-protected Treasury securities; HYG and LQD represent high-yield and investment-grade corporate credit; VIX measures expected equity-market volatility; and VVIX measures the implied volatility of VIX options. For SPY, TIP, HYG, and LQD, the dependent variable is the cumulative abnormal log return over the indicated event window. For VIX and VVIX, it is the cumulative abnormal change in the index level. Solid lines report coefficient estimates, and dashed lines report the corresponding 95% confidence intervals. A positive coefficient indicates that the association between the Human–TTS hawkishness gap and the corresponding market response becomes more positive when economic policy uncertainty is higher. All specifications control for text stance, monetary policy surprises, macroeconomic conditions, the Human–TTS main effect, EPU, the SEP indicator, and chair-regime indicators. HY and IG denote high-yield and investment-grade credit; Vol. and VoV denote equity volatility and volatility of volatility. Orange stars mark estimates with $p < 0.05$.

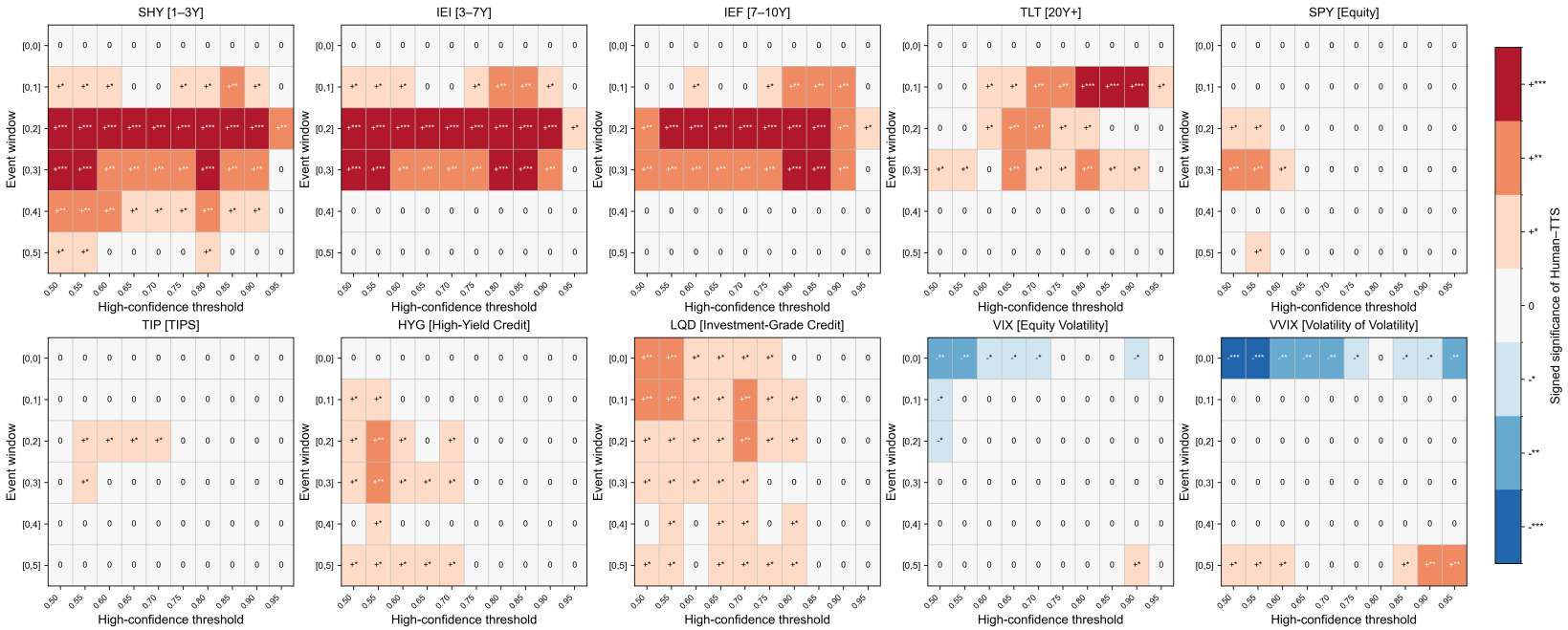


Figure 7: Significance of the Main Human–TTS Net Hawkishness-Gap Effect across High-Confidence Thresholds and Event Windows

Notes: This figure reports the statistical significance of the estimated main effect of the Human–TTS net hawkishness gap across alternative high-confidence thresholds and post-conference event windows. Each cell corresponds to a separate regression specification defined by a particular threshold and event window. Cell shading indicates the statistical significance level of the Human–TTS coefficient, while the sign of the coefficient indicates the direction of the associated market response. Positive coefficients imply that a larger Human–TTS net hawkishness gap is associated with a more positive market response, whereas negative coefficients imply a more negative response. The regressions control for textual hawkishness, monetary policy surprises, macroeconomic conditions, EPU, the $(\text{Human–TTS}) \times EPU$ interaction, the SEP indicator, and chair-regime indicators. Standard errors are clustered by year. ***, **, and * denote $p < 0.01$, $p < 0.05$, and $p < 0.10$, respectively.

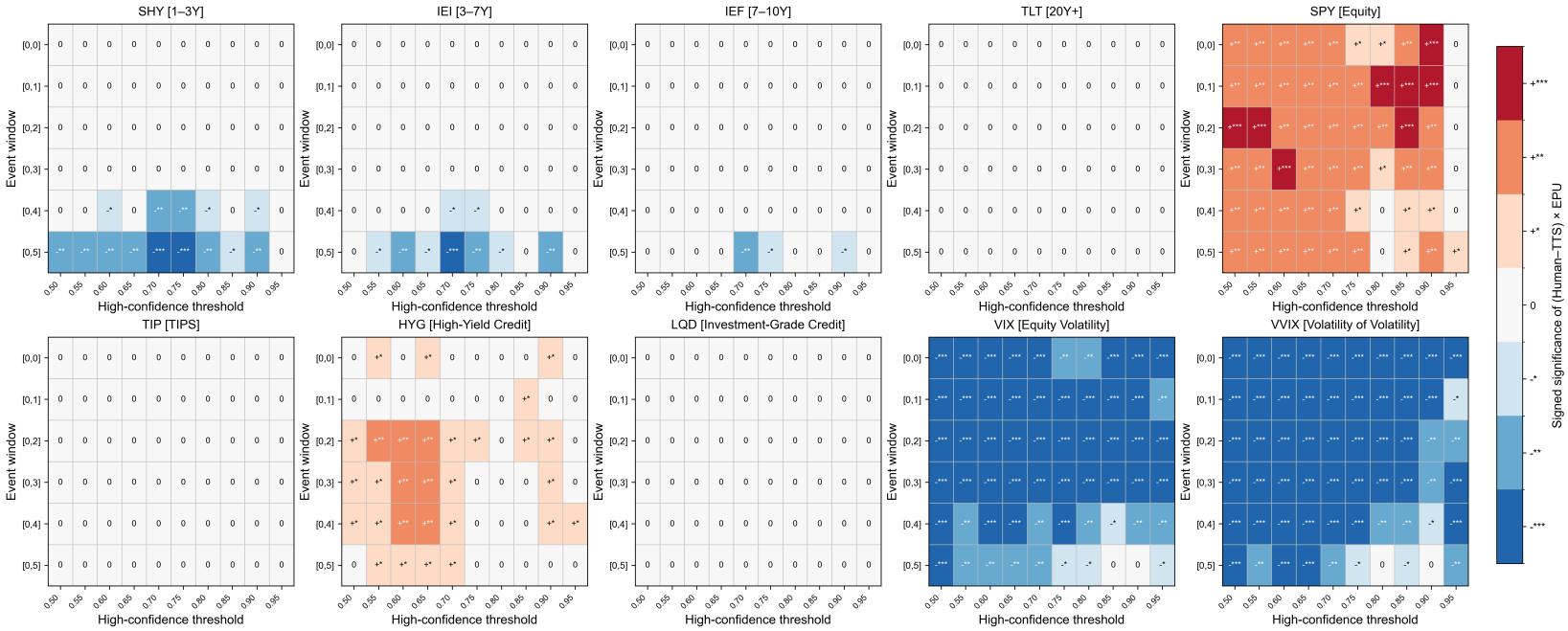


Figure 8: Significance of the Human–TTS × EPU Interaction Effect across High-Confidence Thresholds and Event Windows

Notes: This figure reports the statistical significance of the estimated $(\text{Human–TTS}) \times EPU$ interaction effect across alternative high-confidence thresholds and post-conference event windows. Each cell corresponds to one regression specification defined by a particular high-confidence threshold and event window. Cell shading indicates the statistical significance level of the interaction coefficient, while the sign of the estimated coefficient indicates whether higher economic policy uncertainty strengthens or weakens the market response associated with the Human–TTS net hawkishness gap. Positive coefficients imply that the market response becomes more positive as EPU increases, whereas negative coefficients imply that it becomes more negative. The regressions include the Human–TTS main effect, EPU, textual hawkishness, monetary policy surprises, macroeconomic controls, the SEP indicator, and chair-regime indicators. Standard errors are clustered by year. ***, **, and * denote $p < 0.01$, $p < 0.05$, and $p < 0.10$, respectively.

References

Acosta, M., Ajello, A., Bauer, M. D., Loria, F., and Miranda-Agrippino, S. (2025). Financial market effects of FOMC communication: Evidence from a new event-study database. IMFS Working Paper Series 227, Institute for Monetary and Financial Stability, Goethe University Frankfurt.

Alexopoulos, M., Han, X., Kryvtsov, O., and Zhang, X. (2024). More than words: Fed chairs' communication during congressional testimonies. *Journal of Monetary Economics*, 142:103515.

Aruoba, S. B. and Drechsel, T. (2024). Identifying monetary policy shocks: A natural language approach. NBER Working Paper 32417, National Bureau of Economic Research. Forthcoming, American Economic Journal: Macroeconomics.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636.

Bernanke, B. S. and Kuttner, K. N. (2005). What explains the stock market's reaction to Federal Reserve policy? *The Journal of Finance*, 60(3):1221–1257.

Borochin, P. A., Cicon, J. E., DeLisle, R. J., and Price, S. M. (2018). The effects of conference call tones on market perceptions of value uncertainty. *Journal of Financial Markets*, 40:75–91.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., et al. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Curti, F. and Kazinnik, S. (2023). Let's face it: Quantifying the impact of nonverbal communication in FOMC press conferences. *Journal of Monetary Economics*, 139:110–126.

De Pooter, M. (2021). Questions and answers: The information content of the post-FOMC meeting press conference. FEDS Notes, Board of Governors of the Federal Reserve System. October 12, 2021.

Deng, Y., Xu, M., and Tang, Y. (2024). FMPAF: How do Fed chairs affect the financial market? a fine-grained monetary policy analysis framework on their language. *arXiv preprint arXiv:2403.06115*.

Ewertz, J., Knickrehm, C., Nienhaus, M., and Reichmann, D. (2026). Listen closely: Measuring vocal tone in corporate disclosures. *Journal of Accounting Research*, 64(1):229–277.

Gorodnichenko, Y., Pham, T., and Talavera, O. (2023). The voice of monetary policy. *American Economic Review*, 113(2):548–584.

Gürkaynak, R. S., Sack, B., and Swanson, E. T. (2005). Do actions speak louder than words? the response of asset prices to monetary policy actions and statements. *International Journal of Central Banking*, 1(1):55–93.

Huang, A. H., Wang, H., and Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Jarociński, M. and Karadi, P. (2020). Deconstructing monetary policy surprises—the role of information shocks. *American Economic Journal: Macroeconomics*, 12(2):1–43.

Li, K., Mai, F., Shen, R., and Yan, X. (2021). Measuring corporate culture using machine learning. *The Review of Financial Studies*, 34(7):3265–3315.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lucca, D. O. and Moench, E. (2015). The pre-FOMC announcement drift. *The Journal of Finance*, 70(1):329–371.

Nakamura, E. and Steinsson, J. (2018). High-frequency identification of monetary non-neutrality: The information effect. *The Quarterly Journal of Economics*, 133(3):1283–1330.

Pan, J. and Peng, Q. (2024). The pre-FOMC drift and the secular decline in long-term interest rates. *SSRN Working Paper*. SSRN Abstract ID 4764451.

Pasad, A., Chou, J.-C., and Livescu, K. (2021). Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Shah, A., Paturi, S., and Chava, S. (2023). Trillion dollar words: A new financial dataset, task & market analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6664–6679, Toronto, Canada. Association for Computational Linguistics.

Swanson, E. T. (2021). Measuring the effects of Federal Reserve forward guidance and asset purchases on financial markets. *Journal of Monetary Economics*, 118:32–53.

Appendix for:

Hawkish by Voice: Financial Market Reactions to FOMC Communication

Yueh-Heng Wu[a], Hui-Ching Chuang[a,*], and Ju-Fang Yen[a]

[a]National Taipei University, New Taipei City, Taiwan

[*]Corresponding author

Appendix A. Speech-Model Training and Implementation Details

This appendix documents the pipeline used to construct sentence-level WavLM hawkishness scores: FOMC-RoBERTa pseudo-labeling, audio preprocessing, WavLM-Large fine-tuning, and conference-grouped cross-validation and model selection.

Pseudo-label construction. Fine-tuning WavLM to recognize monetary policy stance from speech requires labeled audio observations. Because manually labeling the full set of sentence-level audio clips would be costly and could introduce inconsistency across annotators, we use FOMC-RoBERTa as a weak-supervision source. In the speech-model pipeline, FOMC-RoBERTa serves as an intermediate labeling device: it converts each sentence's verbal content into a structured policy-stance label, which is then assigned to the corresponding audio.

FOMC-RoBERTa is built on RoBERTa (Liu et al., 2019) and adapted to the FOMC hawkish–dovish–neutral classification task following Shah et al. (2023). This domain-specific classification is important because hawkishness and dovishness refer to monetary policy positions and are not equivalent to generic negative and positive sentiment.

For sentence i, FOMC-RoBERTa produces predicted probabilities $(\widehat{p}_i^H, \widehat{p}_i^D, \widehat{p}_i^N)$, where H, D, and N denote hawkish, dovish, and neutral. The predicted label is $\widehat{y}_i = \arg\max_{c\in\{H,D,N\}} \widehat{p}_i^c$, and its confidence is $q_i = \max_{c\in\{H,D,N\}} \widehat{p}_i^c$. The resulting training sample is $\mathcal{S}_{\text{train}} = \{i : \widehat{y}_i \in \{H, D\},\ q_i \geq 0.80\}$.

Thus, a sentence is retained for fine-tuning only when FOMC-RoBERTa classifies it as hawkish or dovish, and its maximum predicted probability is at least 0.80. The threshold removes sentences

whose policy stance is close to the classification boundary, thereby reducing noise in the supervision labels. It also provides a balance between label reliability and the number of available training observations.

The resulting binary training sample contains 8,256 sentences, comprising 4,209 hawkish sentences and 4,047 dovish sentences. The similarity in the two class sizes reduces the risk that the speech model learns to favor the more frequent class. Neutral and lower-confidence sentences do not enter fine-tuning, but they remain in the full inference sample. Consequently, the trained model can be applied to every valid sentence when constructing sentence- and press-conference-level hawkishness measures. FOMC-RoBERTa's predicted probabilities also provide the text-based benchmark used elsewhere in the empirical analysis.

Audio preprocessing and segmentation. All audio clips are converted to mono waveforms sampled at 16,000 Hz. To accommodate the input-length and GPU-memory requirements of WavLM-Large, we limit each model input to 20 seconds. Sentence clips lasting no more than 20 seconds are used directly.

For clips longer than 20 seconds, we apply a sliding-window procedure with a 20-second window and a two-second overlap. Adjacent windows therefore begin 18 seconds apart. Starting from the beginning of the sentence, the procedure continues until the complete audio clip has been covered. The overlap reduces the possibility that relevant linguistic or acoustic information is lost because it falls exactly at a window boundary.

The same segmentation rule is applied to the high-confidence training sample and the full inference sample. After segmentation, the 8,256 high-confidence training sentences produce 8,687 training windows. The 35,619 cleaned sentences in the full sample produce 36,695 inference windows.

When a sentence is divided into multiple windows, we first obtain a predicted hawkish probability for each window. We then take the arithmetic mean of the window-level probabilities to obtain one sentence-level hawkish probability. The identical aggregation rule is applied to the original human speech and the corresponding TTS audio.

WavLM-Large fine-tuning. WavLM is a self-supervised speech representation model that is first pretrained on large corpora of unlabeled speech and subsequently adapted to a specific downstream task (Chen et al., 2022). Rather than relying only on researcher-selected acoustic variables, the model learns speech representations directly from the waveform.

We fine-tune the pretrained `microsoft/wavlm-large` model for binary hawkish–dovish classification. For audio window s, the model produces logits (z_s^H, z_s^D), which the softmax transformation converts to $\widehat{p}_s^H = \exp(z_s^H)/[\exp(z_s^H) + \exp(z_s^D)]$ and $\widehat{p}_s^D = 1 - \widehat{p}_s^H$.

Our analysis uses $\widehat{p}_s^H$, the predicted hawkish probability, as a continuous measure rather than reducing the output to a binary classification.

Because the available training sample is limited and the labels are generated rather than manually assigned, fully updating the large pretrained model could increase overfitting. We therefore adopt a partial-freezing strategy. We freeze the convolutional feature extractor and the first 18 Transformer encoder layers, and fine-tune only the final six Transformer layers and the classification head.

This design allows the lower layers to retain general acoustic representations learned during pretraining, while the upper layers adapt to the FOMC hawkish–dovish classification task. The approach is also consistent with evidence that different layers of self-supervised speech models encode different combinations of acoustic, phonetic, and higher-level speech information, established for wav2vec 2.0 by Pasad et al. (2021). Attention dropout, hidden-state dropout, and final-layer dropout are each set to 0.1.

Cross-validation and model selection. A conventional random split could place sentences from the same press conference in both the training and validation samples. Such observations share the same conference, chair regime, recording environment, policy context, and background conditions, potentially allowing the model to recognize conference-specific features rather than generalizable policy-stance information.

We therefore use stratified group five-fold cross-validation. The press-conference date defines the group, ensuring that all sentences and windows from a given conference appear in only one fold. Stratification approximately preserves the hawkish–dovish class ratio across folds.

We evaluate the model using three measures. Accuracy is the proportion of correct classifications. Macro F1 calculates the F1 score separately for the hawkish and dovish classes and assigns equal weight to the two classes. The area under the receiver operating characteristic curve (AUC) evaluates the model's ability to distinguish hawkish from dovish speech across alternative classification thresholds.

For each fold, training continues for at most 20 epochs. We retain the checkpoint with the highest validation AUC and stop training when validation performance fails to improve for four consecutive epochs. We pool the out-of-fold predictions from the five validation samples when reporting overall cross-validation performance. Following the implemented model-selection procedure, the fold-specific model with the highest validation AUC is used for subsequent inference on both the original human speech and the corresponding TTS audio.

Table A.1 summarizes the full implementation settings.

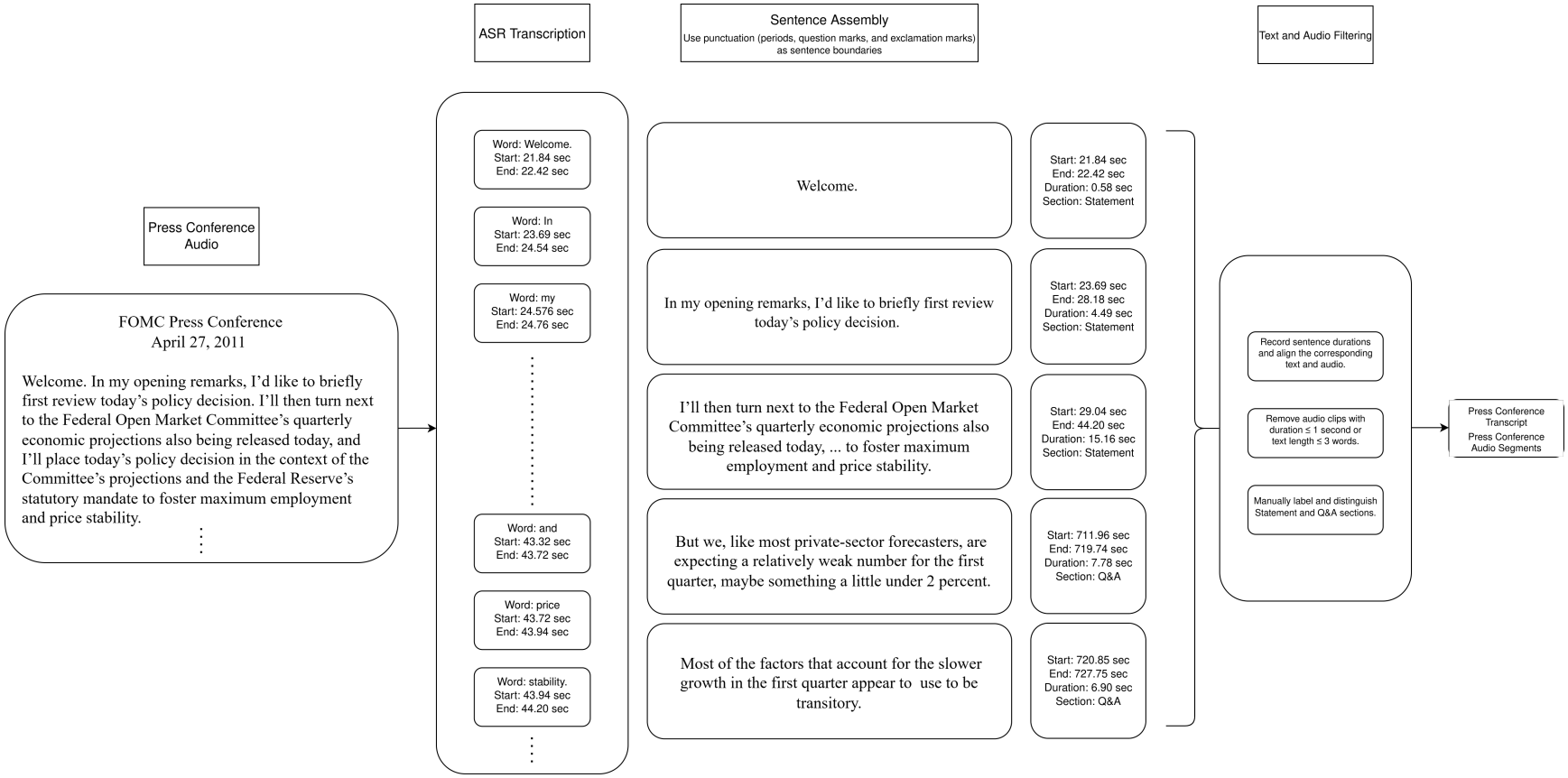


Figure A.1: Automatic speech recognition and filtering process.

Appendix B. Random-Forest and Principal-Component Analysis

This appendix documents the random-forest and principal-component analyses relating eight duration-weighted, meeting-level acoustic features to WavLM-implied hawkishness of the original full-Q&A speech, $H_{m,c}^{(\mathrm{Hum})}$. These exercises are descriptive: unlike the market regressions, which use the high-confidence Human–TTS gap $\Delta S_{m,c}^{\mathrm{Hum-TTS}}(0.80)$, they use the continuous human-speech measure and therefore neither decompose the market coefficients nor isolate the contribution of any speaker role. The index $c \in \{\text{Statement}, \text{Q\&A}\}$ denotes the conference section.

Feature extraction and meeting-level aggregation. Let $x_{i,j}^{(\mathrm{Hum})}$ denote acoustic feature j extracted from the human audio of sentence i, where $j \in \{1, \ldots, 8\}$. The eight features, their abbreviated names, and their interpretations are summarized in Table B.1.

All audio is resampled to 16,000 Hz and converted to a temporary WAV file before feature extraction. Fundamental frequency, or F0, is estimated using Praat-based methods over the range 50–350 Hz, excluding zero values corresponding to unvoiced segments. For the effective F0 sequence $\{f_{i1}, \ldots, f_{iT_i}\}$, the three pitch measures are $\mathrm{MeanF0}_i = T_i^{-1} \sum_{t=1}^{T_i} f_{it}$, $\mathrm{F0STD}_i = [T_i^{-1} \sum_{t=1}^{T_i} (f_{it} - \mathrm{MeanF0}_i)^2]^{1/2}$, and $\mathrm{F0Range}_i = \max_t f_{it} - \min_t f_{it}$. Here, T_i is the number of retained F0 frames for sentence i.

Loudness is calculated from the Praat intensity series after excluding nonpositive values. We define $\mathrm{MeanLoudness}_i$ and $\mathrm{LoudnessSTD}_i$ analogously to Mean F0 and F0 STD. Jitter measures local irregularity in successive vocal-fold periods, shimmer measures local irregularity in successive amplitudes, and HNR measures the strength of periodic harmonic structure relative to noise. Jitter and shimmer are computed using Praat point-process methods, while HNR is obtained from the Praat harmonicity measure.

Sentence-level acoustic features are aggregated to the meeting level using audio-duration weights: $X_{m,c,j}^{(\mathrm{Hum})} = [\sum_{i \in \mathcal{I}_{m,c}} d_i x_{i,j}^{(\mathrm{Hum})}] / [\sum_{i \in \mathcal{I}_{m,c}} d_i]$, where d_i is the aligned duration of sentence i, and $\mathcal{I}_{m,c}$ denotes the set of sentences in conference m and setting c.

Random-forest permutation importance. We use a random-forest regression to examine which meeting-level acoustic features are most predictive of $H_{m,c}^{(\text{Hum})}$. The model allows for nonlinearities and interactions among the eight acoustic features without imposing a parametric functional form.

Predictive performance is evaluated using five-fold cross-validation. Within each validation fold, permutation importance is calculated by randomly permuting one feature while leaving the remaining features unchanged. Let $R_{f,0}^2$ denote the validation R^2 in fold f before permutation, and let $R_{f,j,r}^2$ denote the validation R^2 after feature j is permuted in repetition r. The associated decline is $\Delta R_{f,j,r}^2 = R_{f,0}^2 - R_{f,j,r}^2$. Averaging over $F = 5$ folds and $R = 100$ repetitions gives $I_{j,c}^{(\text{perm})} = (FR)^{-1} \sum_{f=1}^{F} \sum_{r=1}^{R} \Delta R_{f,j,r}^2$. A larger value of $I_{j,c}^{(\text{perm})}$ indicates that the feature provides greater incremental predictive information. Values near zero indicate little contribution, while negative values indicate unstable predictive importance. The error bars in Figure B.1 report the standard deviation of $\Delta R_{f,j,r}^2$ across folds and repetitions.

Table B.2 summarizes the random-forest implementation.

Principal-component analysis and regression. Because the eight acoustic features are measured on different scales and may be strongly correlated, we use principal-component analysis to summarize them into a smaller number of orthogonal acoustic dimensions. For setting c, each meeting-level feature is standardized as $X_{m,c,j}^* = (X_{m,c,j}^{(\text{Hum})} - \overline{X}_{c,j})/s_{c,j}$, where $\overline{X}_{c,j}$ and $s_{c,j}$ are its sample mean and standard deviation. Stacking the standardized observations into $\mathbf{X}_c^*$ yields the score matrix $\mathbf{P}_c = \mathbf{X}_c^* \mathbf{V}_c$, where $\mathbf{V}_c = (\mathbf{v}_{1,c}, \ldots, \mathbf{v}_{8,c})$ contains the eigenvectors and $\lambda_{q,c}$ denotes the eigenvalue of component q. The corresponding loading is $L_{j,q,c} = v_{j,q,c}\sqrt{\lambda_{q,c}}$, which incorporates both the feature's weight and the variance represented by the component.

The principal-component regression is $H_{m,c}^{(\text{Hum})} = \alpha_c + \beta_{1,c}\text{PC}_{m,c,1} + \beta_{2,c}\text{PC}_{m,c,2} + \beta_{3,c}\text{PC}_{m,c,3} + \varepsilon_{m,c}$. The model is estimated by OLS, and the figures report coefficient estimates, 95% confidence intervals, and statistical significance.

Q&A acoustic interpretation. Figure B.1 reports random-forest permutation importance for the Q&A. Jitter produces the largest average decline in validation R^2, followed by mean pitch and loudness variation. The remaining features make smaller contributions. Because the error bars

are relatively wide and some intervals cross zero, the ranking should be interpreted as descriptive rather than as a precise ordering of effect sizes. The results nevertheless indicate that human hawkishness is jointly associated with voice perturbation, pitch, and loudness dynamics rather than being dominated by a single acoustic feature.

Figure B.2 shows that the first three principal components explain more than 80% of the variation in the eight meeting-level acoustic features. We therefore focus on these three components in the loading and regression analyses.

Figure B.3 reports the principal-component loadings. PC1 contrasts voice clarity with vocal perturbation: it loads positively on HNR, mean pitch, and loudness variation and negatively on jitter, shimmer, and pitch variation. Higher PC1 values therefore correspond to clearer and more stable speech, whereas lower values correspond to greater frequency and amplitude perturbation.

PC2 primarily represents pitch variation, with its largest loadings on F0 Range and F0 STD. PC3 captures vocal force and loudness control: it loads positively on mean loudness and negatively on loudness variation and shimmer. Higher PC3 values therefore correspond to stronger average vocal intensity, more stable loudness, and lower amplitude perturbation.

Figure B.4 reports the principal-component regression estimates. PC1 is negatively and statistically significantly associated with human hawkishness, whereas PC3 is positively and significantly associated with human hawkishness. PC2 is not statistically significant.

The negative PC1 coefficient implies that higher WavLM-implied hawkishness is associated with the lower-PC1 acoustic profile: greater jitter and shimmer, lower HNR, and greater vocal perturbation. This pattern is consistent with greater vocal tension or on-the-spot pressure during real-time exchanges. The positive PC3 coefficient indicates that higher human hawkishness is also associated with greater mean loudness, lower loudness variation, and lower shimmer, a profile consistent with more forceful and controlled delivery.

These patterns are not mutually exclusive. An interactive Q&A conveying a clearer restrictive stance may simultaneously display greater real-time vocal tension and stronger control over vocal intensity. The interpretation remains suggestive: the estimates do not establish a media-

tion mechanism linking individual acoustic features, the Human–TTS gap, and financial-market responses.

The random-forest and PCA evidence complements the main-text Human–TTS comparisons by showing that WavLM-implied vocal hawkishness is systematically associated with measurable acoustic variation. These results should be interpreted as model interpretation rather than as a causal decomposition of the voice gap or the market-response coefficients.

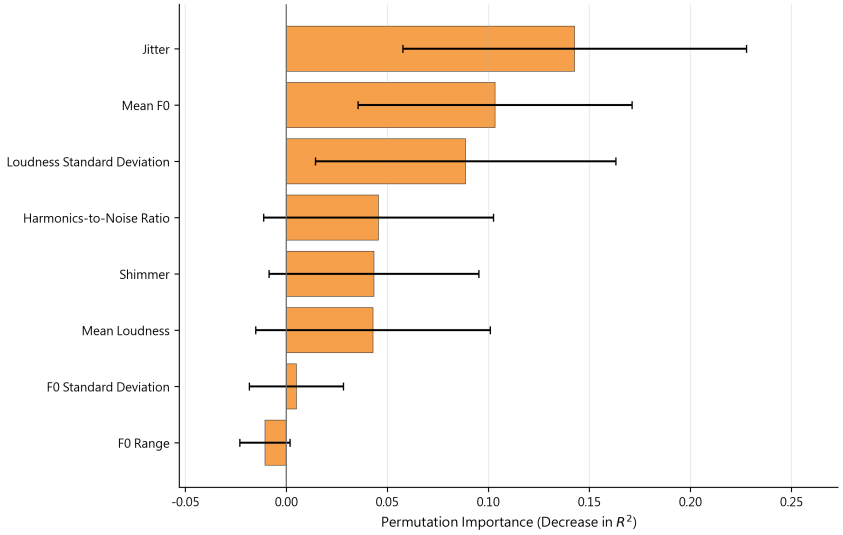


Figure B.1: Random-Forest Permutation Importance of Acoustic Features

Notes: This figure reports permutation importance for eight meeting-level acoustic features used to predict WavLM-implied hawkishness of the original full-Q&A speech. Importance is measured by the decline in out-of-sample R^2 when a feature is randomly permuted in a validation fold. Larger values indicate a greater contribution to predictive performance. Error bars report the standard deviation across five folds and 100 permutation repetitions per fold.

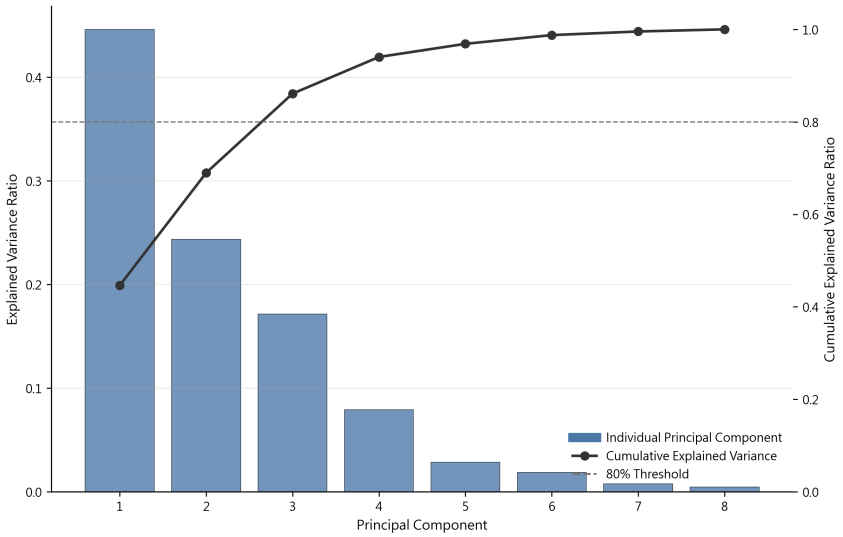


Figure B.2: Variance Explained by the Q&A Acoustic Principal Components

Notes: The bars report the share of variance explained by each principal component, and the line reports cumulative explained variance. PCA is applied to the eight standardized meeting-level human acoustic features during the Q&A.

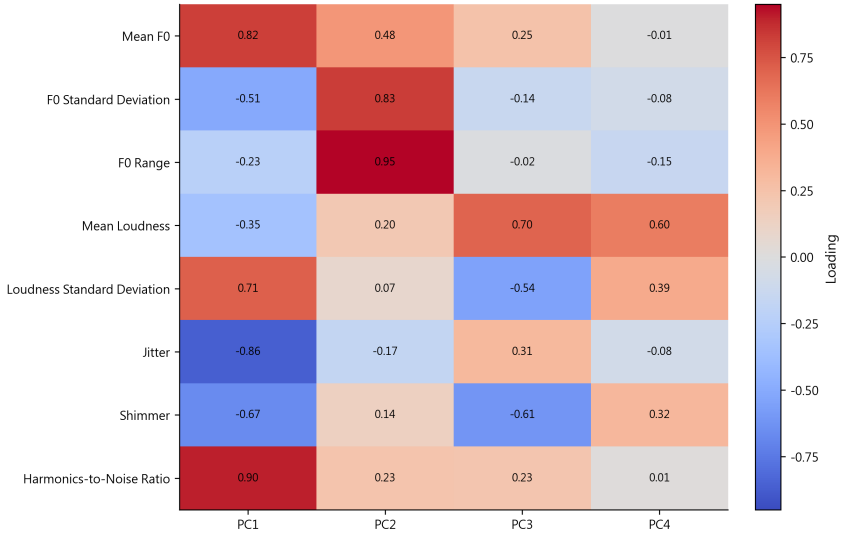


Figure B.3: Principal-Component Loadings of Q&A Acoustic Features

Notes: This figure reports the loadings of the eight standardized meeting-level acoustic features on the principal components constructed from the full Q&A speech. Positive and negative values indicate the direction of each feature's contribution to a component.

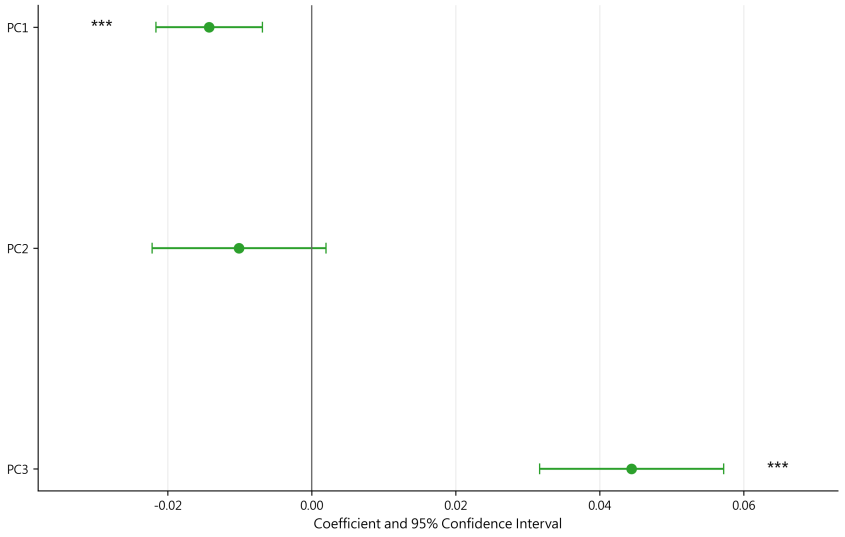


Figure B.4: Principal-Component Regression of Human Hawkishness during the Q&A

Notes: This figure reports OLS coefficients from a regression of meeting-level WavLM-implied human hawkishness on the first three principal components of the standardized Q&A acoustic features. Error bars report 95% confidence intervals. ***, **, and * denote $p < 0.01$, $p < 0.05$, and $p < 0.10$, respectively.

Table A.1: Speech-Model Training and Implementation Settings

Item	Setting
Panel A: Pseudo-labels and audio preparation	
Supervision model	FOMC-RoBERTa
Predicted text classes	Hawkish, dovish, and neutral
Speech-model training classes	Hawkish and dovish
Confidence requirement	Maximum predicted class probability ≥ 0.80
High-confidence training sample	8,256 sentences: 4,209 hawkish and 4,047 dovish
Audio format	16,000 Hz mono waveform
Maximum input length	20 seconds
Sliding-window overlap	2 seconds, corresponding to an 18-second stride
Training windows	8,687
Full-sample inference windows	36,695
Multiple-window aggregation	Arithmetic mean of window-level hawkish probabilities
Panel B: Model architecture	
Base model	`microsoft/wavlm-large`
Classification task	Binary hawkish–dovish classification
Frozen components	Convolutional feature extractor and first 18 Transformer layers
Trainable components	Final six Transformer layers and classification head
Model output	Continuous hawkish probability obtained by softmax
Attention dropout	0.10
Hidden-state dropout	0.10
Final-layer dropout	0.10
Panel C: Estimation and validation	
Learning rate	3×10^{-5}
Maximum epochs	20
Training batch size per GPU	8
Gradient accumulation steps	4
Warmup ratio	0.10
Learning-rate scheduler	Cosine
Weight decay	0.05
Label smoothing	0.10
Numerical precision	FP16 mixed-precision training
Early-stopping patience	4 epochs
Cross-validation	Stratified group five-fold cross-validation
Grouping variable	Press-conference date
Performance measures	Accuracy, macro F1, and AUC
Within-fold checkpoint criterion	Highest validation AUC
Model used for subsequent inference	Model from the fold with the highest validation AUC

Table B.1: Definitions of Acoustic Features

Short name	Acoustic feature	Definition
Mean F0	Mean pitch	Average fundamental frequency of the voice.
F0 STD	Pitch STD	Variation in fundamental frequency within a sentence.
F0 Range	Pitch range	Difference between the maximum and minimum fundamental frequencies.
Mean Loudness	Mean loudness	Average vocal intensity.
Loudness STD	Loudness STD	Variation in vocal intensity within a sentence.
Jitter	Jitter	Cycle-to-cycle irregularity in vocal-fold frequency.
Shimmer	Shimmer	Cycle-to-cycle irregularity in vocal-fold amplitude.
HNR	Harmonics-to-noise ratio	Strength of periodic harmonic energy relative to noise; higher values generally indicate a clearer and more periodic voice signal.

Table B.2: Main Random-Forest Settings

Item	Setting
Unit of analysis	Meeting level
Outcome	$H_{m,c}^{(\mathrm{Hum})}$
Predictors	Eight human acoustic features
Cross-validation	Five-fold
Number of trees	1,000
Minimum samples per leaf	3
Minimum samples to split	4
Candidate features per split	`sqrt`
Bootstrap sampling	Yes
Permutation repetitions per fold	100
Importance measure	Validation R^2 decline

Notes: Permutation importance is computed on the validation observations in each fold. Larger declines in validation R^2 indicate greater predictive importance.

Table B.3: $CAR[0,2]$ Regressions for Treasury ETFs

	SHY [1–3Y]	IEI [3–7Y]	IEF [7–10Y]	TLT [20Y+]
Human–TTS	0.008***	0.018***	0.027***	0.058*
	(3.439)	(3.098)	(3.097)	(1.923)
Text	-0.004	-0.005	-0.006	0.011
	(-1.183)	(-0.560)	(-0.463)	(0.409)
MP1	-0.038	-0.121	-0.204	-0.432
	(-1.099)	(-1.470)	(-1.462)	(-1.372)
MP2	-0.042***	-0.084***	-0.123**	-0.161
	(-4.240)	(-3.179)	(-2.564)	(-1.219)
Unemp	-0.000	0.000	0.000	0.001
	(-0.051)	(0.644)	(0.637)	(0.713)
PCE	-0.000	-0.000	-0.001	-0.003***
	(-0.942)	(-0.985)	(-1.225)	(-2.630)
EPU	-0.000	-0.000	-0.000	-0.000
	(-0.965)	(-1.242)	(-1.434)	(-1.501)
$(\text{Human–TTS}) \times EPU$	-0.000	-0.000	-0.000	0.000
	(-0.567)	(-0.376)	(-0.346)	(0.063)
SEP	-0.001	-0.002*	-0.003	-0.002
	(-1.165)	(-1.903)	(-1.458)	(-0.495)
Bernanke	0.001	0.000	-0.000	-0.003
	(1.438)	(0.267)	(-0.106)	(-0.669)
Yellen	0.001	0.003	0.004	0.007
	(1.243)	(1.146)	(1.194)	(0.958)
Constant	-0.003	-0.002	-0.002	0.011
	(-0.911)	(-0.293)	(-0.212)	(0.477)
Observations	89	89	89	89
R^2	0.426	0.347	0.317	0.285
Adjusted R^2	0.344	0.254	0.219	0.183

Notes: This table re-estimates Equation (5) over the $[0,2]$ event window and is the robustness counterpart to Table 9, which uses $[0,1]$. The dependent variable is the cumulative abnormal log return over the event window $[0,2]$, denoted $CAR_{m,a,[0,2]}$. SHY, IEI, IEF, and TLT track Treasury securities with maturities of 1–3, 3–7, 7–10, and more than 20 years, respectively. Human–TTS is the high-confidence net hawkishness of all original full-Q&A speech—chair answers, reporter questions, and moderator remarks—minus that of the same-text TTS audio. Text is the corresponding high-confidence net hawkishness constructed from verbal content. MP1 and MP2 measure the current policy-rate surprise and the revision in the expected policy rate at the next scheduled FOMC meeting. Unemp and PCE are mean-centered unemployment and year-over-year PCE inflation; EPU is the mean-centered Economic Policy Uncertainty index. SEP indicates the release of a Summary of Economic Projections. Bernanke and Yellen are chair-regime indicators, with the Powell regime omitted. EPU is measured in index points, so its coefficients and those of the interaction term are small by construction; entries shown as 0.000 are nonzero but smaller than 0.0005 in absolute value. Year-clustered t-statistics are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table B.4: $CAR[0,2]$ Regressions for Equity, Inflation, Credit, and Volatility Markets

	SPY [Equity]	TIP [TIPS]	HYG [HY]	LQD [IG]	VIX [Vol.]	VVIX [VoV]
Human–TTS	0.013	0.026	0.034	0.058*	-4.151	9.684
	(0.311)	(1.425)	(1.189)	(1.656)	(-0.724)	(1.023)
Text	0.026	0.004	0.005	0.005	-4.019	-2.858
	(0.715)	(0.280)	(0.295)	(0.280)	(-0.688)	(-0.206)
MP1	0.072	-0.284**	-0.009	-0.241	-48.659	-132.371
	(0.247)	(-2.349)	(-0.045)	(-1.100)	(-1.216)	(-0.859)
MP2	-0.177	-0.118***	-0.142***	-0.178**	9.692	-23.075
	(-1.220)	(-3.231)	(-2.706)	(-2.361)	(0.448)	(-0.342)
Unemp	-0.001	0.001	0.001	0.002	0.096	1.239***
	(-0.442)	(0.786)	(1.239)	(1.057)	(0.278)	(4.972)
PCE	0.001	-0.001	0.000	-0.001	-0.403	-1.182
	(0.311)	(-1.006)	(0.091)	(-0.492)	(-0.991)	(-1.510)
EPU	-0.000	-0.000*	-0.000	-0.000*	0.002	-0.003
	(-0.736)	(-1.699)	(-1.394)	(-1.712)	(0.757)	(-0.479)
(Human–TTS) $\times$ EPU	0.000**	0.000	0.000	0.000	-0.060***	-0.131***
	(2.221)	(1.183)	(1.604)	(1.067)	(-3.050)	(-2.753)
SEP	-0.016**	-0.004**	-0.008**	-0.005**	0.720	-0.180
	(-2.410)	(-2.071)	(-2.417)	(-2.022)	(0.732)	(-0.071)
Bernanke	0.018***	0.001	0.006	0.006	-3.383***	-5.890***
	(3.265)	(0.224)	(1.003)	(0.967)	(-4.778)	(-3.116)
Yellen	0.017	0.003	0.012	0.011	-3.580**	-5.666
	(1.437)	(0.637)	(1.549)	(1.281)	(-2.033)	(-1.611)
Constant	0.020	0.005	0.004	0.004	-2.204	-0.975
	(0.638)	(0.404)	(0.303)	(0.249)	(-0.419)	(-0.077)
Observations	89	89	89	89	89	89
R^2	0.206	0.271	0.238	0.355	0.274	0.180
Adjusted R^2	0.093	0.167	0.129	0.263	0.170	0.063

Notes: This table re-estimates Equation (5) over the $[0,2]$ event window and is the robustness counterpart to Table 10, which uses $[0,1]$. The dependent variable is $CAR_{m,a,[0,2]}$. For SPY, TIP, HYG, and LQD, CAR is a cumulative abnormal log return; for VIX and VVIX, it is a cumulative abnormal change in the index level, so the coefficients in those two columns are not comparable in scale with the return columns. Human–TTS is the high-confidence net hawkishness of all original full-Q&A speech—chair answers, reporter questions, and moderator remarks—minus that of the same-text TTS audio. Text is high-confidence net textual hawkishness. MP1 and MP2 are the current-rate and expected-path surprises. Unemp, PCE, and EPU are mean-centered. SEP is the Summary of Economic Projections indicator. Bernanke and Yellen are chair-regime indicators, with the Powell regime omitted. EPU is measured in index points, so its coefficients and those of the interaction term are small by construction; entries shown as 0.000 are nonzero but smaller than 0.0005 in absolute value. Year-clustered t-statistics are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels.