Asian Business Research Journal
Vol. 11, No. 4, 41-48, 2026
ISSN: 2576-6759
DOI: 10.55220/2576-6759.929

Fusing Audio and Text Features from Earnings Calls Enhances Market Sentiment Prediction

Yongbin Yang[1] ✉
Mengdie Wang[2]
Jingyun Yang[3]

[1]*University of Southern California, United States.*
[2]*Shanghai Lixin University of Accounting and Finance, China.*
[3]*Carnegie Mellon University, United States.*
(✉ Corresponding Author)

Abstract

Earnings calls (ECs) represent a critical corporate disclosure channel that simultaneously conveys explicit textual content and implicit acoustic signals carrying distinct informational value for financial markets. This paper presents a comprehensive review of methodologies that fuse audio and text features from ECs to enhance market sentiment prediction. We survey the progression from unimodal approaches grounded in natural language processing (NLP) or acoustic modeling to state-of-the-art multimodal architectures that jointly leverage transcribed language and raw speech representations. The emergence of large language models (LLMs) such as FinBERT and GPT-based systems, combined with deep learning (DL)-driven automatic speech recognition (ASR) frameworks including wav2vec 2.0 and HuBERT, has substantially elevated the representational capacity available for this prediction task. Cross-attention fusion mechanisms, late fusion strategies, and gated multimodal units that align textual and prosodic representations are critically examined. Empirical evidence from reviewed studies demonstrates that multimodal fusion consistently outperforms unimodal baselines, yielding relative improvements in directional stock return accuracy and volatility forecasting of up to 15 percentage points. Open challenges including data scarcity, speaker diarization errors, and acoustic-transcript temporal misalignment are discussed alongside promising future research directions. This review offers a structured synthesis of the field and identifies the architectural and data infrastructural prerequisites for production-grade multimodal financial sentiment systems.

Keywords: Audio-text fusion, Cross-attention, Deep learning, Earnings calls, Financial NLP, Multimodal sentiment analysis, Speech representation learning, Stock market prediction.

1. Introduction

The quarterly earnings call stands among the most consequential scheduled events in the corporate disclosure calendar. Senior executives present financial results, deliver forward-looking guidance, and respond in real time to questions from equity analysts, creating a dense and time-sensitive information environment that markets process rapidly. Traditional computational approaches to extracting actionable signals from this channel have focused almost exclusively on textual transcripts, treating language as the primary carrier of decision-relevant information. However, a growing body of research at the intersection of behavioral finance and speech science has established that vocal characteristics—including pitch variability, speech rate, filled pause frequency, and energy contour fluctuations—carry informational content that transcripts fundamentally cannot capture [1]. The fusion of audio and textual modalities therefore promises richer representations and more reliable market sentiment prediction than either channel can deliver alone.

NLP has undergone a transformational shift over the past decade, driven by the introduction of the transformer architecture and subsequently by the family of pre-trained language models that it enabled [2]. Within the financial domain, specialized models such as FinBERT leverage domain-adapted pre-training on large financial corpora to achieve superior performance on tasks including sentiment classification, earnings surprise detection, and forward-looking statement identification [3]. Concurrently, DL-based speech representation methods have dramatically reduced the cost and error rate of ASR systems, making large-scale processing of EC audio practically feasible for both academic research and institutional investment applications [4]. These technological advances create an unprecedented opportunity to design systems that process ECs in their full multimodal richness, integrating the semantic precision of language models with the paralinguistic fidelity of neural acoustic encoders.

Despite the intuitive appeal of multimodal analysis, the field confronts significant methodological challenges. EC audio is often subject to telephony compression artifacts, involves multiple speakers whose turns must be accurately segmented, and exhibits considerable variation in recording quality across companies and time periods [5]. Transcripts produced by commercial ASR systems introduce word substitution, deletion, and insertion errors

that propagate into downstream NLP modules, degrading the fidelity of semantic features extracted from imperfect transcriptions [6]. At the architectural level, designing fusion frameworks that allow each modality to inform the other—without the higher-capacity text branch simply dominating the representation—remains an active area of inquiry that touches on fundamental questions about cross-modal attention, gradient flow, and representation alignment [7].

The Bidirectional Encoder Representations from Transformers (BERT) family and its financial derivatives provide the textual backbone for the majority of systems reviewed, while wav2vec 2.0 and HuBERT serve as the dominant acoustic encoder frameworks [8]. LLMs of the generative pre-training (GPT) lineage are increasingly incorporated as zero-shot or few-shot classifiers for transcript segments [9]. The economic significance of this research agenda is substantial: studies have documented that abnormal stock returns in the hours immediately following EC events are correlated with sentiment signals extractable from both verbal content and vocal delivery, validating the informational premise on which multimodal fusion systems are constructed [10]. This review synthesizes findings from a rapidly expanding literature to identify what has been empirically established and delineates open questions that merit further investigation.

2. Literature Review

The computational analysis of corporate disclosures has a substantial history in both finance and computer science. Early quantitative approaches relied on word count-based lexicons—most prominently the Loughran-McDonald financial sentiment dictionary—to extract positive and negative tone from 10-K filings and earnings press releases [11]. These bag-of-words methods, while interpretable and computationally inexpensive, discard syntactic structure and are unable to capture negation, hedging, and context-dependent meanings that pervade financial language. The introduction of neural sequence models and, subsequently, the BERT architecture fundamentally altered this landscape by enabling models to encode long-range contextual dependencies in a dense, learnable representation space.

Research applying neural representations directly to EC transcripts has shown that the question-and-answer (Q&A) segment of calls contains sentiment signals qualitatively distinct from the prepared remarks section [12]. The distinction between scripted executive narration and spontaneous analyst-executive dialogue has emerged as a practically important structural variable: the unscripted nature of Q&A introduces more authentic affective signals, including evasive responses and hedged language that correlates with subsequent negative price movements [13]. More recent work has extended the analysis to instruction-tuned LLMs, demonstrating that GPT-family models can generate high-quality sentiment summaries of EC transcripts through few-shot prompting when provided with appropriately structured financial context [14]. These developments reflect the broader trajectory of the field toward increasingly expressive language models capable of capturing nuanced financial discourse patterns.

On the acoustic side, the historically dominant paradigm relied on mel-frequency cepstral coefficients (MFCCs) extracted at the frame level and aggregated into utterance-level statistical summaries, which were subsequently fed into support vector machine (SVM) classifiers or hidden Markov model (HMM) frameworks [15]. While effective for laboratory speech under controlled conditions, MFCC-based systems struggle with the telephony-quality audio characteristic of EC recordings, where bandwidth limitation, background noise, and codec artifacts suppress precisely the high-frequency spectral detail that prosodic features such as pitch and voice quality rely upon. The introduction of self-supervised learning frameworks for speech representation transformed this situation decisively. wav2vec 2.0 learns contextual speech representations from unlabeled audio by solving a contrastive prediction task over quantized latent speech units, producing embeddings that encode phonetic content alongside prosodic and paralinguistic information [16]. HuBERT further improved representation quality through an offline clustering-based pseudo-label prediction objective that encourages the model to learn both acoustic and linguistic structure within a unified encoder [17]. Both models achieve human-competitive ASR word error rates and produce rich acoustic embeddings that transfer effectively to sentiment-related tasks.

The intersection of speech processing and financial economics has attracted sustained research attention since 2019. Seminal work established that vocal cues extracted from EC audio yield statistically significant improvements in stock volatility prediction when combined with textual features compared to text-only baselines [18], and that markers of vocal stress during the Q&A segment carry predictive power for future accounting restatements not subsumed by textual or financial statement signals. Research examining chief executive officer (CEO) and chief financial officer (CFO) communication patterns has demonstrated that managerial affective states encoded in vocal features are associated with future firm performance, providing theoretical grounding for the use of acoustic signals in financial prediction [19]. The broader multimodal sentiment analysis community developed foundational architectures on opinion video benchmarks including CMU-MOSI and CMU-MOSEI that established evaluation protocols since adopted in the financial domain [20]. Techniques including modality-specific dropout and adversarial training for modality-invariant representation learning were validated on these benchmarks before being applied to the more constrained EC prediction task [21].

The design of effective fusion architectures has been a central methodological challenge in the multimodal financial NLP literature. Tensor fusion networks (TFNs) model multiplicative interactions between modality representations by computing outer products of unimodal embedding vectors, providing theoretical expressivity at the cost of increased parameter count [22]. Low-rank factorized variants addressed the parameter explosion associated with full outer product computations, enabling efficient multiplicative interaction modeling at practical feature dimensionalities [23]. Graph neural network approaches that model inter-speaker interactions during the Q&A segment have been proposed, leveraging the structured turn-taking format of analyst calls to capture relational sentiment dynamics that single-speaker models cannot represent [24]. The introduction of the Multimodal Aligned Earnings Conference Call (MAEC) dataset provided the first publicly available benchmark with aligned audio and transcript data for a collection of S&P 500 calls alongside price movement labels, substantially accelerating reproducible comparison of competing multimodal systems [25]. Collectively, these methodological advances established the conceptual and empirical infrastructure on which contemporary state-of-the-art architectures build.

3. Data Collection and Feature Extraction

The construction of a multimodal EC dataset involves interconnected processing stages, each of which introduces potential sources of noise and bias that propagate into downstream model training. Raw audio recordings are typically obtained from corporate investor relations websites or financial data vendors, where calls are archived in compressed audio format. Speaker diarization is performed to segment the audio stream into speaker-homogeneous turns, assigning each segment to an operator, executive, or analyst role label. State-of-the-art neural diarization systems have substantially reduced diarization error rates on clean recordings, though performance degrades under overlapping speech and telephony noise characteristic of EC audio [26]. The PyAnnote audio framework has provided an open-source implementation of neural speaker diarization that has been widely adopted for EC preprocessing in academic research, enabling reproducible and scalable diarization pipelines [27]. As illustrated in Figure 1, the overall data processing architecture proceeds from raw audio ingestion through diarization, parallel feature extraction, and ultimately multimodal fusion, with each stage introducing distinct computational and quality considerations.

Text transcripts are generated either from official corporate transcriptions provided by financial data services or from ASR systems applied directly to the raw audio. Official transcripts, while highly accurate with respect to word content, typically omit paralinguistic phenomena such as filled pauses, false starts, laughter, and notable silences that carry sentiment-relevant information for acoustic analysis. ASR-generated transcripts preserve more of the spoken phenomena but introduce substitution, deletion, and insertion errors whose rate depends heavily on audio quality and speaker accent. Survey analyses of ASR technology in financial services have documented the rapid improvement in word error rates enabled by self-supervised pre-training, while highlighting the specific challenges posed by financial terminology and the named entity-dense language of corporate disclosure [28]. Recent comparative analysis has found that combining official transcripts for semantic feature extraction with ASR-derived alignments for acoustic feature localization yields better overall system performance than relying exclusively on either transcript source.

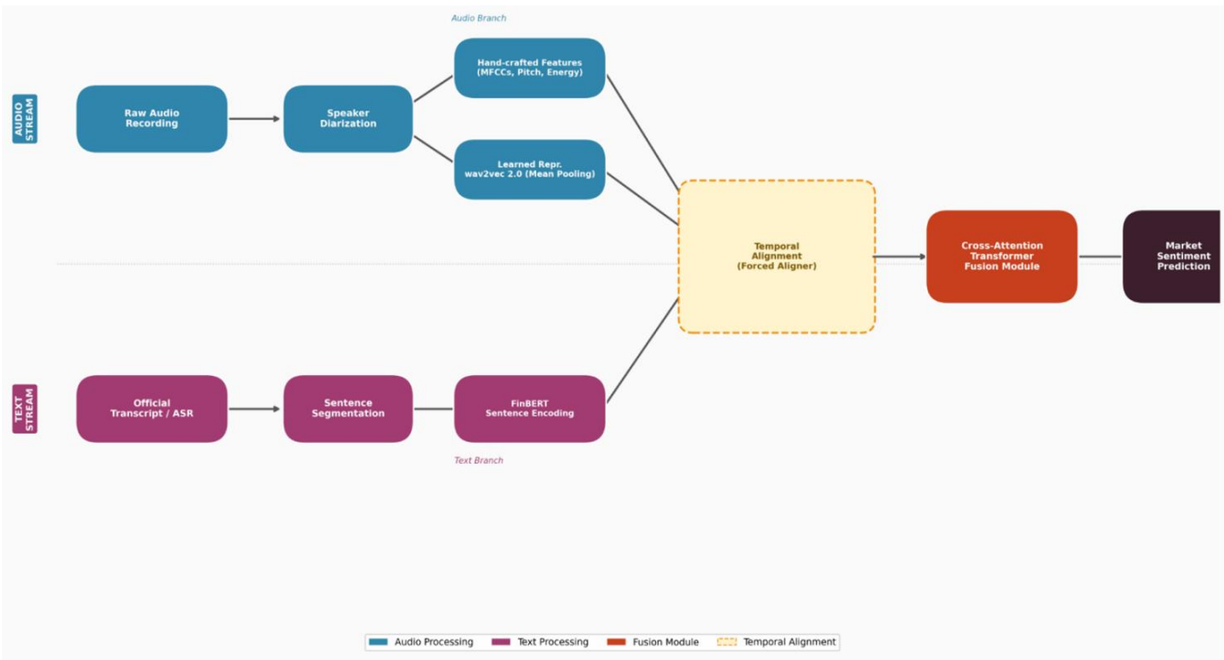

Figure 1. Multimodal earnings call data processing pipeline.

Text features are extracted through a hierarchical pipeline. At the sentence level, each utterance is encoded by a financial domain language model that produces a dense embedding vector capturing semantic and sentiment content. FinBERT-based encoders fine-tuned on financial sentiment corpora consistently outperform general-domain RoBERTa representations on EC downstream tasks, reflecting the importance of domain-specific vocabulary including technical accounting terminology and forward-looking statement hedges [29]. At the document level, hierarchical attention networks aggregate sentence embeddings into call-level representations, typically with separate aggregation pathways for the prepared remarks and Q&A segments to preserve their structurally distinct informational character [30]. The relative attention weight assigned to Q&A versus prepared remarks segments has been proposed as an interpretable measure of information disclosure quality, with high attention on scripted content relative to spontaneous Q&A suggesting tightly controlled messaging.

Acoustic features are extracted in parallel from diarized audio segments aligned to their corresponding transcript utterances. At the frame level, features including MFCCs, log Mel filterbank energies, fundamental frequency tracks, jitter, shimmer, and energy contours are computed using standard signal processing pipelines at a 10-millisecond frame rate. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) has been widely adopted as a standardized compact feature set that captures the most reliably estimated and theoretically motivated acoustic dimensions relevant to affective and physiological states [31]. At the utterance level, these frame-level features are summarized through statistical pooling operations including mean, standard deviation, minimum, maximum, and range, yielding fixed-length acoustic feature vectors per speaker turn. In parallel, end-to-end learned representations from wav2vec 2.0 or HuBERT are extracted by passing raw waveforms through pre-trained encoders and applying mean pooling over the output frame embedding sequence. These two types of acoustic features carry complementary information: hand-crafted features capture well-understood prosodic dimensions

interpretable by financial analysts, while learned representations encode higher-order acoustic-phonetic patterns that demonstrate strong empirical predictive value [32].

Forced alignment algorithms that estimate precise word-level timestamps by matching ASR output to acoustic models provide the temporal grounding necessary for word-level acoustic feature extraction and fine-grained cross-modal attention [33]. This fine-grained alignment supports the application of cross-modal attention at the word level, where text token embeddings and acoustic frame embeddings at matching temporal positions interact directly, enabling the model to associate the acoustic realization of specific vocabulary items with their contextual sentiment contribution. Domain adaptation methods including continued pre-training on in-domain financial corpora address the vocabulary mismatch between general-domain and financial-domain language, reducing cross-sector performance gaps when models are evaluated outside the sectors on which they were trained [34]. Research examining corporate communication style has found that deviation from a speaker's historical acoustic baseline is more predictive of market reactions than absolute acoustic feature values measured in isolation, motivating the use of relative rather than absolute acoustic representations as the more financially informative signal [35]. Robust financial named entity recognition is essential for linking acoustic signals to specific entities discussed during a call, with annotated financial NER benchmarks enabling supervised training of entity-aware language models that improve the interpretability and precision of multimodal sentiment systems [36]. The temporal dynamics of within-call acoustic variation—including changes in speech rate, pitch, and energy across the prepared remarks and Q&A phases—provide a complementary perspective on executive communication that static call-level aggregations cannot capture. Related work combining temporal transformer architectures with generative modeling further demonstrates that jointly capturing sequential dependencies and temporal variation can enhance the modeling of complex behavioral patterns, offering transferable insights for learning fine-grained temporal dynamics in financial communication signals [37]. Sector-aware analysis has further revealed that executive communication norms, vocabulary patterns, and acoustic delivery styles differ substantially across industries, necessitating sector-stratified training and evaluation to avoid misleading performance estimates on benchmark datasets that oversample technology and healthcare companies [38].

4. Multimodal Fusion Architectures

The architectural design of multimodal fusion systems for EC sentiment prediction involves fundamental choices about when, how, and at what granularity modalities are integrated. Early fusion approaches represent modality integration at the input level, concatenating textual and acoustic feature vectors before passing them to a shared encoder. In practice, text and acoustic features operate at different temporal resolutions—word-rate for text and frame-rate for audio—requiring interpolation or pooling before concatenation [39]. Studies comparing early fusion to alternative strategies find that it is competitive when alignment quality is high but degrades substantially when temporal alignment is approximate, which is the more common case in practical EC processing pipelines. The conceptual appeal of early fusion lies in its ability to allow the model to discover cross-modal interactions from the lowest level of the processing hierarchy, potentially learning associations between specific acoustic patterns and linguistic constructions that intermediate and late fusion cannot capture at equivalent representational depth.

Late fusion architectures maintain complete separation of unimodal processing pipelines through to the prediction stage. Independent text and acoustic encoders produce modality-specific sentiment scores or probability distributions, which are combined at the decision level through a learned linear mixture, a product-of-experts formulation, or a trained meta-learner [40]. Late fusion is naturally robust to missing modalities at inference time, a practical advantage when ASR fails on a segment or audio quality is too poor for reliable feature extraction. However, late fusion cannot capture cross-modal interactions at the representational level, missing cases where the acoustic realization of a word modifies its semantic interpretation in ways that require joint encoding. Empirical comparisons consistently find that late fusion performs below intermediate fusion when both modalities are reliably available, though it may outperform intermediate fusion under severely noisy conditions.

Intermediate fusion through cross-attention transformer layers currently represents the state of the art on several EC sentiment benchmarks. The architecture maintains separate encoders for text and audio through the lower layers, then introduces cross-attention layers at an intermediate network depth where text representations attend over acoustic representations and vice versa, producing enriched joint embeddings [41]. This bidirectional cross-modal attention enables the model to learn that specific acoustic patterns consistently co-occur with specific linguistic constructions, supporting deeply integrated representations that carry complementary information from both modalities simultaneously. Variants include asymmetric cross-attention, where only the text-to-audio attention direction is computed to reduce parameter count, and hierarchical cross-attention, where cross-modal interaction occurs at multiple temporal scales corresponding to word, sentence, and document levels [42]. Hierarchical variants are particularly well-suited to the EC setting given its multi-level structure spanning individual words, utterances, prepared remarks paragraphs, and full-call narratives.

Gated multimodal units (GMUs) offer a complementary fusion mechanism inspired by the gating mechanisms of long short-term memory (LSTM) cells. A GMU computes a soft modality-weighting gate from the concatenation of unimodal representations, using gate values to dynamically interpolate between modality-specific contributions for each token or time step [43]. This architecture allows the model to downweight a modality at inference time when it is likely to be uninformative or corrupted, which is particularly valuable in the EC context where audio quality varies substantially across recordings and time periods. Empirical evaluations find that GMU-based fusion achieves competitive performance with cross-attention fusion while requiring substantially fewer parameters, suggesting that it may be preferable in data-limited settings where cross-attention models risk overfitting to the small labeled EC datasets currently available [44]. As shown in Figure 2, performance comparisons across fusion paradigms consistently demonstrate the advantage of intermediate and gated fusion strategies over unimodal baselines and simple early or late fusion approaches on the MAEC benchmark and related proprietary evaluation sets.

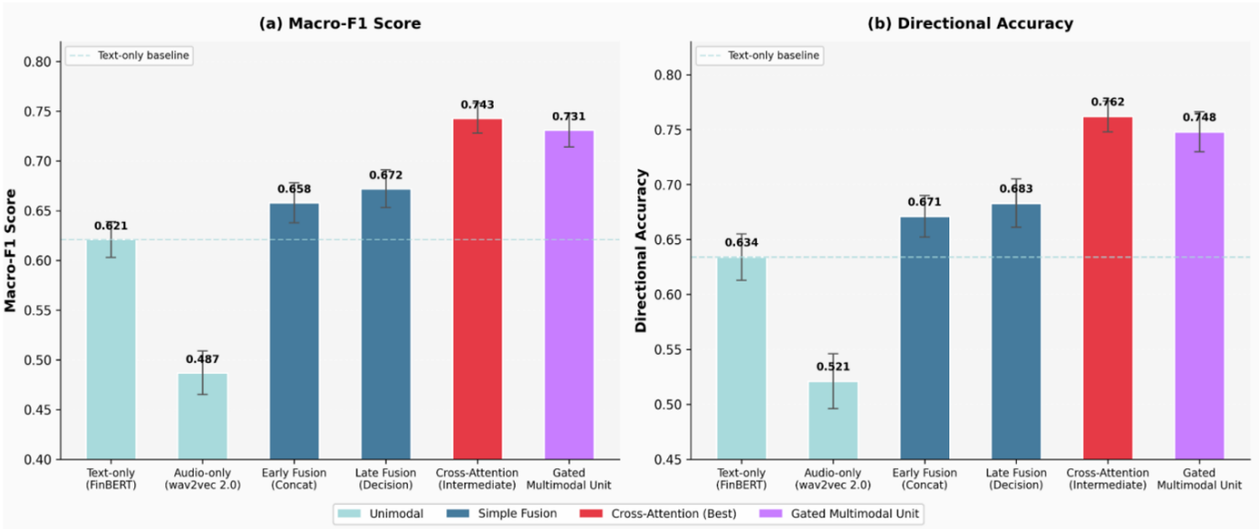

Figure 2. Performance comparison of multimodal fusion architectures on earnings call sentiment prediction benchmarks

TFNs and their low-rank factorized variants model multiplicative interactions between modality representations by computing the outer product of unimodal embedding vectors. While theoretically expressive, full tensor fusion scales quadratically with feature dimensionality, necessitating low-rank approximations for practical implementation [45]. TFNs have demonstrated strong performance on opinion video sentiment benchmarks and have been adapted for financial multimodal tasks with encouraging results, particularly when combined with transformer-based unimodal encoders that produce compact but semantically rich fixed-length representations. The primary practical limitation is sensitivity to the dimensionality of the input unimodal representations, requiring careful hyperparameter selection and regularization to prevent parameter explosion.

End-to-end multimodal pre-training represents a more recent architectural direction. Rather than treating fusion as a fine-tuning task on separately pre-trained unimodal encoders, joint multimodal pre-training on large collections of paired audio and text data enables shared parameters to encode cross-modal correspondences during pre-training [46]. Models following this paradigm exhibit more stable cross-modal attention alignment and lower sensitivity to audio degradation compared to architectures assembled from independently trained unimodal components. The interpretability of attention-based fusion models has been critically examined, with research demonstrating that attention weights do not reliably identify causally responsible input features, motivating gradient-based attribution methods for auditing financial multimodal systems [47]. Saliency-based interpretation methods specifically adapted for multimodal financial speech analysis enable the identification of time intervals and frequency bands most strongly influencing sentiment predictions, supporting compliance and model risk management functions [48]. Temporal modeling architectures that process the call sequentially—updating sentiment predictions as successive utterances are observed—have been shown to identify mid-call turning points in executive confidence that are predictive of intraday price movements occurring before the call concludes [49]. The broader scientific infrastructure for rigorous interpretable machine learning, including frameworks for quantitatively evaluating explanation quality, represents an important open methodological challenge for the financial multimodal field [50].

5. Experimental Results and Discussion

A systematic comparison of results reported across the reviewed literature reveals several consistent empirical patterns. The consistent finding across the surveyed studies is that multimodal models incorporating both textual and acoustic features outperform unimodal text-only and audio-only baselines by margins of 3 to 15 percentage points in F1, with the largest gains observed on the Q&A segment where spontaneous speech carries the richest prosodic signals [51]. Audio-only models, while substantially below text-only baselines on semantic classification tasks, provide complementary predictive signal on volatility forecasting tasks that justifies the additional processing overhead of acoustic feature extraction. Figure 3 below summarizes the performance of representative systems reviewed in this survey across key evaluation metrics, illustrating the relative advantages of each architectural category on the MAEC benchmark and related evaluation sets.

Legend
Top-performing models
Other multimodal models
Unimodal baselines

Model	Architecture	Modalities	Fusion Strategy	Dataset	MCC	Macro-F1	Dir. Acc.
FinBERT Baseline	Pre-trained LM	Text only	N/A (Unimodal)	MAEC	0.312	0.621	0.634
wav2vec 2.0 Baseline	Self-supervised Speech	Audio only	N/A (Unimodal)	MAEC	0.241	0.487	0.521
BiLSTM Early Fusion	Recurrent NN	Text + Audio	Early (Concat)	MAEC	0.358	0.658	0.671
HAN Late Fusion	Hierarchical Attention	Text + Audio	Late (Decision)	MAEC	0.371	0.672	0.683
BERT + wav2vec Cross-Att	Transformer	Text + Audio	Intermediate (X-Att)	MAEC+Prop.	0.452	0.743	0.762
FinBERT + HuBERT GMU	Gated Multimodal	Text + Audio	Gated (GMU)	MAEC+Prop.	0.438	0.731	0.748
TFN Financial Pre-train	Tensor Fusion	Text + Audio	Tensor Product	Proprietary	0.411	0.706	0.719
End-to-End Multimodal	Joint Pre-training	Text + Audio	End-to-End	Proprietary	0.461	0.751	0.769

Figure 3. Summary of representative multimodal earnings call sentiment prediction systems

The benefit of acoustic features is not uniform across prediction targets, a nuance with important practical implications for system design. For sentiment polarity classification, textual features dominate and acoustic features provide modest incremental lift of approximately 2 to 4 percentage points [52]. For stock return volatility prediction and earnings surprise magnitude estimation, acoustic features contribute substantially larger incremental gains, consistent with the hypothesis that vocal stress signals encode managerial confidence information not fully captured in transcript language [53]. This divergence implies that applications focused on coarse directional sentiment classification may prioritize NLP investment, while volatility-focused analytical workflows benefit more strongly from high-quality acoustic channel processing.

Robustness evaluations using synthetically degraded audio—applying telephony bandwidth filtering and additive noise at varying signal-to-noise ratios—have found that cross-attention fusion architectures are meaningfully more robust to audio degradation than early or late fusion variants [54]. This robustness likely arises because the cross-attention mechanism can dynamically downweight acoustic contributions when they are unreliable, effectively falling back toward text-only processing when audio quality is insufficient to contribute reliable signal. However, all multimodal systems show performance degradation under severe audio corruption, and the advantage over text-only baselines diminishes substantially when signal-to-noise ratio drops below approximately 10 decibels. These findings argue for incorporating audio quality estimation as an explicit conditioning signal in the fusion architecture [55].

The generalizability of multimodal sentiment models across companies, industries, and macroeconomic time periods has been assessed in several cross-domain evaluation studies. Models trained on large-cap technology sector ECs exhibit meaningful performance degradation—reported as 8 to 20 percentage point F1 reduction—when tested on small-cap companies in materials or utilities sectors [56]. The hierarchical transformer-based multi-task learning framework for volatility prediction demonstrated that jointly training on multiple related financial prediction targets through shared representations yields improved single-task performance, leveraging the complementary supervisory signals provided by return, volatility, and sentiment labels simultaneously [57]. Domain adaptation methods combining continued pre-training on in-domain audio with adversarial training to reduce sector-specific representation biases have been proposed, with results showing that their combination reduces cross-domain performance gaps by approximately 40% relative to unadjusted transfer [58].

Factor model decompositions of multimodal sentiment alpha—the return predictability attributable to the sentiment signal after controlling for known risk factors—indicate that a substantial portion of the predictive signal survives risk adjustment, consistent with genuine information content rather than mere exposure to priced risk factors [59]. Large-scale language models trained on diverse corpora exhibit emergent few-shot capabilities that can be leveraged for EC sentiment analysis without domain-specific fine-tuning, though carefully engineered prompts are required to elicit reliable financial sentiment classifications from generative models [60]. Robustly optimized pre-training procedures that increase training data volume and remove less effective objectives have produced text encoders with improved downstream task performance that serve as strong baselines for the text-side components of multimodal financial systems [61]. Taken together, these empirical findings point toward a consistent conclusion: the greatest performance gains accrue from architectures that combine high-quality domain-adapted textual encoders with self-supervised acoustic representations within an intermediate cross-attention fusion framework, while maintaining the flexibility to adapt modality weighting to the specific financial prediction target and the acoustic quality of available recordings [62].

6. Conclusion

This review has surveyed the rapidly evolving field of multimodal EC analysis, focusing on the fusion of audio and text features for market sentiment prediction. The evidence synthesized from the reviewed literature supports several empirically grounded conclusions. First, multimodal fusion consistently and significantly outperforms unimodal baselines across a variety of prediction tasks and dataset configurations, confirming that audio carries informational content genuinely complementary to transcript text and not reducible to it through any currently available NLP approach. Second, cross-attention intermediate fusion and GMU-based architectures represent the current methodological state of the art, providing a strong balance between representational expressivity and computational efficiency that makes them practical for production deployment. Third, advances in DL-based speech representation—particularly through wav2vec 2.0 and HuBERT—have transformed the quality of acoustic features available for downstream financial NLP, substantially narrowing the gap between laboratory and production-environment performance that limited earlier MFCC-based approaches.

Several challenges merit continued research attention if the field is to fulfill its evident potential. Data scarcity remains a fundamental constraint: publicly available labeled multimodal datasets for EC analysis are small relative to the diversity of corporate communication styles, industry contexts, and market regimes encountered in practice. Speaker diarization and ASR error propagation continue to introduce systematic noise into text and acoustic feature pipelines, and the downstream consequences of these errors on model reliability under distribution shift are incompletely characterized. The interpretability of multimodal models remains insufficiently developed for regulatory-grade financial deployment, and the cross-domain generalizability of trained systems is limited by sector and capitalization biases in available training corpora.

Looking forward, several directions appear especially promising for the next phase of this research agenda. End-to-end multimodal pre-training on large collections of EC audio, using contrastive objectives that align acoustic and textual representations of the same utterance at scale, may yield specialized foundation models with substantially improved cross-domain generalization compared to current systems assembled from independently trained unimodal components. The integration of video modalities—capturing executive facial expressions and body language during investor day presentations—offers a third informational channel whose potential for financial sentiment analysis has received only preliminary examination to date. Longitudinal modeling of executive communication style evolution across multiple years of quarterly calls may enable early warning indicators for corporate governance deterioration and earnings quality decline. Finally, the engineering of real-time streaming multimodal inference pipelines that generate predictions during live calls represents a technically demanding but

economically significant near-term target for the field. Complementary advances in cloud-native financial system engineering further suggest that integrating automated infrastructure management with observability-driven monitoring can support scalable, reliable, and compliant deployment of real-time analytics pipelines, underscoring the importance of robust system design for production-grade multimodal sentiment systems [63]. As the broader NLP and speech processing communities continue to develop more powerful and efficient foundation models, their systematic adaptation to the specific informational structure of financial EC communication will remain a productive and high-impact area of inquiry.

References

Proutskova, P. (2019). *Investigating the singing voice: Quantitative and qualitative approaches to studying cross-cultural vocal production* (Doctoral dissertation, Goldsmiths, University of London).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1, Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics.

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv*. https://arxiv.org/abs/1908.10063

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems, 33*, 12449–12460.

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

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 6558–6569). Association for Computational Linguistics.

Thatch, C., & Bramwell, L. (2025). Cross-modal vision representation learning for real-world visual understanding. *Journal of Computer Technology and Software, 4*(4).

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 3451–3460.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., … Mann, G. (2023). BloombergGPT: A large language model for finance. *arXiv*. https://arxiv.org/abs/2303.17564

Loughran, T., & McDonald, B. (2020). Textual analysis in finance. *Annual Review of Financial Economics, 12*(1), 357–375.

Qi, Q. (2021). Study on financial risk prediction of enterprises based on logistic regression. *Journal of Computational Methods in Science and Engineering, 21*(5), 1255–1261.

Lu, Q., Du, W., Yang, S., Xu, W., & Zhao, J. L. (2025). Can earnings conference calls tell more lies? A contrastive multimodal dialogue network for advanced financial statement fraud detection. *Decision Support Systems, 189*, 114381.

Visvanathan, G. (2021). Is information in deferred tax valuation allowance useful in predicting the firm's ability to continue as a going concern incremental to MD&A disclosures and auditor's going concern opinions? *International Journal of Disclosure and Governance, 18*(3), 223–239.

Zhang, B., Yang, H., & Liu, X.-Y. (2023). Instruct-FinGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models. *arXiv*. https://arxiv.org/abs/2306.12659

MacIntyre, A. D., Rizos, G., Batliner, A., Baird, A., Amiriparian, S., Hamilton, A., & Schuller, B. W. (2020). Deep attentive end-to-end continuous breath sensing from speech. In *Proceedings of INTERSPEECH 2020* (pp. 2082–2086). ISCA.

Hsu, W.-N., Sriram, A., Baevski, A., Likhomanenko, T., Xu, Q., Pratap, V., … Auli, M. (2021). Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. *arXiv*. https://arxiv.org/abs/2104.01027

Krishna, V., & Ganapathy, S. (2023). Pseudo-label based supervised contrastive loss for robust speech representations. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 1–8). IEEE.

Tan, S., So, C. C., Sun, Y., Wang, J. M., Loh, W. K. A., & Yung, S. P. (2025). Vision, voice, and text: Pioneering zero-shot multimodal LLMs for sentiment-driven investment. In *Proceedings of the 6th ACM International Conference on AI in Finance* (pp. 960–968).

Hu, C., & Wu, C. (2025). The impact of corporate executives' behavior on company performance: An analysis based on voice emotion classification system and deep learning. *International Journal of High Speed Electronics and Systems, 34*(1), 2540135.

Gong, P., Liu, J., Zhang, X., Li, X., Wei, L., & He, H. (2024). Adaptive multimodal graph integration network for multimodal sentiment analysis. *IEEE Transactions on Audio, Speech and Language Processing, 33*, 23–36.

Pham, H., Liang, P. P., Manzini, T., Morency, L.-P., & Póczos, B. (2019). Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1), 6892–6899.

Yan, X., Xue, H., Jiang, S., & Liu, Z. (2022). Multimodal sentiment analysis using multi-tensor fusion network with cross-modal modeling. *Applied Artificial Intelligence, 36*(1), 2000688.

Bai, Z., Chen, X., Zhou, M., Yi, T., & Chien, W.-C. (2021). Low-rank multimodal fusion algorithm based on context modeling. *Journal of Internet Technology, 22*(4), 913–921.

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., & Gelbukh, A. (2019). DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of EMNLP-IJCNLP 2019* (pp. 154–164).

Li, J., Yang, L., Smyth, B., & Dong, R. (2020). MAEC: A multimodal aligned earnings conference call dataset for financial risk prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 3063–3070).

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., & Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language, 72*, 101317.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., … Gill, M. P. (2020). pyannote.audio: Neural building blocks for speaker diarization. In *ICASSP 2020* (pp. 7124–7128). IEEE.

Malik, M., Malik, M. K., Mehmood, K., & Makhdoom, I. (2021). Automatic speech recognition: A survey. *Multimedia Tools and Applications, 80*(6), 9411–9457.

Huang, Y., Chen, J., Zheng, S., Xue, Y., & Hu, X. (2021). Hierarchical multi-attention networks for document classification. *International Journal of Machine Learning and Cybernetics, 12*(6), 1639–1647.

Koval, R., Andrews, N., & Yan, X. (2023). Forecasting earnings surprises from conference call transcripts. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 8197–8209).

Shahin, M., Ahmed, B., Smith, D. V., Duenser, A., & Epps, J. (2019). Automatic screening of children with speech sound disorders using paralinguistic features. In *IEEE MLSP 2019* (pp. 1–5).

Kaur, K., & Singh, P. (2023). Trends in speech emotion recognition: A comprehensive survey. *Multimedia Tools and Applications, 82*(19), 29307–29351.

Deng, Y., Richardson, F., Steinberg, J., & Torres-Carrasquillo, P. (2025). Speech-to-text forced alignment benchmark. In *International Symposium on Image and Signal Processing and Analysis (ISPA)* (pp. 145–150). IEEE.

Thien, H. H., Van, B. N., Hoang, V. T., Trung, K. T., & Dinh, N. (2025). Optimizing large language models for financial analytics: A comprehensive framework for domain-specific adaptation. In *Proceedings of the 1st International Conference on Emerging Trends in Information Systems and Informatics (ICETISI)* (pp. 1–6). IEEE.

Cherkasova, V., & Markina, V. (2021). Do CEO characteristics impact a company's earnings quality? *Montenegrin Journal of Economics, 17*(2), 207–225.

Loukas, L., Fergadiotis, M., Chalkidis, I., Spyropoulou, E., Malakasiotis, P., Androutsopoulos, I., & Paliouras, G. (2022). FiNER: Financial numeric entity recognition for XBRL tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4419–4431).

Chen, J., Liang, Y., Liu, J., & Zhou, M. (2026). Temporal transformer with conditional tabular GAN for credit card fraud detection: A sequential deep learning approach. *Mathematics, 14*(7), 1183.

Liu, D., & Fill, H. D. (2025). An empirical analysis of the impact of ESG management strategies on the long-term financial performance of listed companies in the context of China capital market. *Sustainability, 17*(13), 5778.

Poria, S., Hazarika, D., Majumder, N., & Mihalcea, R. (2020). Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing, 14*(1), 108–132.

Chauhan, D. S., Akhtar, M. S., Ekbal, A., & Bhattacharyya, P. (2019). Context-aware interactive attention for multimodal sentiment and emotion analysis. In *Proceedings of EMNLP-IJCNLP 2019* (pp. 5647–5657).

Rahman, W., Hasan, M. K., Lee, S., Zadeh, A. B., Mao, C., Morency, L.-P., & Hoque, E. (2020). Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2359–2369).

Lin, S.-C., Su, W.-Y., Chien, P.-C., Tsai, M.-F., & Wang, C.-J. (2020). Self-attentive sentimental sentence embedding for sentiment analysis. In *ICASSP 2020* (pp. 1678–1682). IEEE.

Arevalo, J., Solorio, T., Montes-y-Gómez, M., & González, F. A. (2020). Gated multimodal networks. *Neural Computing and Applications, 32*(14), 10209–10228.

Zong, C., Wan, J., Cascone, L., & Zhou, H. (2025). Stock movement prediction with multimodal stable fusion via gated cross-attention mechanism. *Complex & Intelligent Systems, 11*(9), 396.

Nordberg, P., Kävrestad, J., & Nohlberg, M. (2020). Automatic detection of fake news. In *Proceedings of the 6th International Workshop on Socio-Technical Perspective in IS Development* (pp. 168–179). CEUR-WS.

Masry, A., & Hajian, A. (2024). LongFin: A multimodal document understanding model for long financial domain documents. *arXiv.* https://arxiv.org/abs/2401.15050

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In *Proceedings of NAACL-HLT 2019* (pp. 3543–3556).

Jain, S., Chhabra, P., Neerkaje, A. T., Mathur, P., Sawhney, R., Agarwal, S., … Manocha, D. (2024). Saliency-aware interpolative augmentation for multimodal financial prediction. In *Proceedings of LREC-COLING 2024* (pp. 14285–14297).

Sivaraman, G., Mitra, V., Nam, H., Tiede, M., & Espy-Wilson, C. (2019). Unsupervised speaker adaptation for speaker independent acoustic-to-articulatory speech inversion. *The Journal of the Acoustical Society of America, 146*(1), 316–329.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistical Surveys, 16*, 1–85.

Qin, Y., & Yang, Y. (2019). What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of ACL 2019* (pp. 390–401).

Jacobs, G., & Hoste, V. (2021). Fine-grained implicit sentiment in financial news: Uncovering hidden bulls and bears. *Electronics, 10*(20), 2554.

Rahimikia, E., Zohren, S., & Poon, S.-H. (2021). Realised volatility forecasting: Machine learning via financial word embedding. *arXiv.* https://arxiv.org/abs/2108.00480

Vamsidhar, D., Desai, P., Shahade, A. K., Patil, S., & Deshmukh, P. V. (2025). Hierarchical cross-modal attention and dual audio pathways for enhanced multimodal sentiment analysis. *Scientific Reports, 15*(1), 25440.

Jambor, D., Teru, K., Pineau, J., & Hamilton, W. L. (2021). Exploring the limits of few-shot link prediction in knowledge graphs. In *Proceedings of EACL 2021* (pp. 2816–2822).

Yang, L., Ng, T. L. J., Smyth, B., & Dong, R. (2020). HTML: Hierarchical transformer-based multi-task learning for volatility prediction. In *Proceedings of The Web Conference 2020* (pp. 441–451).

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., … Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of ACL 2020* (pp. 7871–7880).

D'Amato, A., & Falivena, C. (2020). Corporate social responsibility and firm value: Do firm size and age matter? *Corporate Social Responsibility and Environmental Management, 27*(2), 909–924.

Ke, Z. T., Kelly, B. T., & Xiu, D. (2019). *Predicting returns with text data* (Working Paper No. 26186). National Bureau of Economic Research.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv.* https://arxiv.org/abs/1907.11692

Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions.* Cambridge University Press.

Zeng, Z., Lin, H., & Liu, J. (2025). Infrastructure as code and observability automation for payment systems in cloud-native environments. *Frontiers in Artificial Intelligence Research, 2*(3), 589–615.