CAN SPEECH FOUNDATION MODELS REPRESENT MEANING THROUGH PROSODY? EVALUATING PRODUCTION AND PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern speech foundation models can process and generate fluent speech, but spoken meaning is not determined by words alone. We ask whether these models can use prosody to convey and recover meaning when lexical content is held constant. Using syntactically ambiguous sentences whose interpretations can be disambiguated by prosodic phrasing, we evaluate the same communicative contrasts bidirectionally. In perception, speech-capable models infer intended interpretations from human productions; in production, models generate the same sentences for specified interpretations, and target-blind human listeners report the meanings they recover. Human listeners recover intended meanings reliably, whereas current speech models perform substantially worse and exhibit strong interpretation and response-position biases. Diagnostic analyses reveal distinct failure modes: models may respond to prosodic variation without reliably grounding their analyses in the acoustic signal, recover informative acoustic cues without mapping them consistently to linguistic structure, or verbalize appropriate prosodic knowledge without using it functionally. In production, measurable acoustic differentiation and naturalness likewise do not guarantee successful communication, while realizations more aligned with human phrasing tend to support better meaning recovery. Together, these results show that representing prosody requires more than preserving acoustic information or producing prosodic variation: acoustic realization, prosodic organization, linguistic structure, and communicative meaning must be functionally aligned.

1 INTRODUCTION

Speech foundation models increasingly process and generate speech directly, but access to the acoustic signal does not by itself imply that a model represents how acoustic variation contributes to linguistic meaning. A system may preserve timing, pitch, or intensity information, or even verbalize appropriate knowledge about prosody, without using that information reliably in interpretation. More generally, information being present or decodable does not establish that it plays a functional role in model behavior (Hewitt & Liang, 2019). This raises a broader representation-learning question: *how are representations at different levels of abstraction connected so that perceptual variation can constrain linguistic interpretation?*

Prosody provides a particularly revealing test of this problem. Prosodic functions are realized through continuous, multidimensional, and context-dependent acoustic patterns rather than fixed forms (Pierrehumbert & Hirschberg, 1990; Ladd, 2008; Wagner & Watson, 2010). A pause, lengthening, or pitch movement is not intrinsically associated with one meaning; its interpretation depends on where it occurs and how it relates to surrounding linguistic structure. Prior work has shown that prosodic phrasing reflects systematic interactions among acoustic cues, syntactic constituency, and prominence, and that the mapping between prosodic and syntactic phrasing is systematic but not one-to-one (Kuang et al., 2022b;a; 2023). Human listeners nevertheless use prosodic information rapidly during sentence processing (Steinhauer et al., 1999). Successful prosodic interpretation therefore requires more than encoding acoustic detail: variable acoustic evidence must be organized into prosodic structure and related to higher-level linguistic representations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Structural ambiguity makes this mapping unusually tractable. The same sequence of words can support multiple syntactic interpretations, while prosodic phrasing can favor one analysis over another (Price et al., 1991; Snedeker & Trueswell, 2003; Kuang et al., 2022b). For example, *David cleaned the mirror in the hallway* is lexically compatible with more than one attachment structure, but different phrasings can bias the listener toward different interpretations. Because lexical content is held constant, the contrast cannot be resolved from the transcript alone. This provides a controlled test of a specific representational interface:

acoustic realization $\leftrightarrow$ prosodic structure $\leftrightarrow$ syntactic structure $\leftrightarrow$ communicative interpretation.

A model may succeed at some links while failing at others: it may detect an acoustic difference without identifying its prosodic significance, know abstractly that phrase boundaries constrain syntax without grounding that knowledge in the current waveform, or infer a plausible prosodic structure while allowing a response preference to dominate the final decision.

We therefore treat prosodic representation as a set of related but empirically distinguishable competencies. **Acoustic sensitivity** asks whether variation in the speech signal affects behavior. **Acoustic grounding** asks whether observed acoustic evidence supports the prosodic structure attributed to the utterance. **Prosodic–linguistic mapping** asks whether that structure is related appropriately to the relevant syntactic representation. Finally, **functional use** asks whether these representations jointly support the intended communicative interpretation. These distinctions are analytical rather than a claim that humans or models implement a fixed sequence of discrete processing stages.

The distinction between acoustic form and communicative function is also important for speech generation. Recent studies have shown that contemporary TTS systems vary substantially in how accurately they realize prosodic information structure, and that naturalness does not necessarily imply successful communication of the intended prosodic function (Chan & Kuang, 2024; Kuang et al., 2026; Chan & Kuang, 2025). The same issue arises for phrasing: a system may generate measurable timing or pitch differences between intended meanings without placing those differences in human-consistent locations or making them interpretable to a listener.

We evaluate this representational mapping in both directions. In **perception**, speech models hear human productions of lexically identical ambiguous sentences and infer the intended structural interpretation. We go beyond aggregate accuracy by asking whether matched changes in prosody actually change model judgments, whether those judgments are robust to answer-order reversal, and whether the prosodic structures reported in model explanations align with independently measured acoustic evidence. In **production**, models generate the same lexical sentence for different intended interpretations. We assess whether those meanings produce systematic acoustic differences, whether the differences resemble human meaning-dependent phrasing patterns, and, most importantly, whether target-blind human listeners recover the intended interpretation.

Human speech serves as a functional reference rather than a fixed acoustic template. Human speakers realize the same structural distinction through variable combinations of cues, while listeners integrate those cues with linguistic structure to recover relatively stable interpretations (Kuang et al., 2022a; 2023). We therefore define successful representation in terms of *functional alignment*: whether acoustic realization, prosodic organization, linguistic structure, and communicative interpretation are connected in a way that preserves meaning from speaker to listener.

Our contributions are threefold:

1. We introduce a linguistically controlled test of **cross-representational alignment**, holding lexical content constant while prosodic phrasing changes the intended syntactic interpretation.
2. We distinguish **acoustic sensitivity, acoustic grounding, prosodic–linguistic mapping, and functional use**, and use complementary behavioral, acoustic, and explanation-based diagnostics to identify where these mappings succeed or fail.
3. We evaluate prosodic representation **bidirectionally**: models must both recover meaning from human prosody and realize intended meaning in generated prosody that independent human listeners can recover.

Together, these analyses ask not simply whether speech models represent acoustics, prosody, or syntax individually, but whether those representations are *functionally aligned*. Prosody thus provides a

controlled case for a broader problem in foundation models: how representations at different levels of abstraction interact to support meaningful behavior.

2 Related Work

Prosody as an interface between acoustic and linguistic structure. Prosodic interpretation requires relating continuous acoustic evidence to higher-level linguistic organization. In syntactic disambiguation, listeners use phrasing cues such as pauses, preboundary lengthening, and tonal structure to constrain competing parses (Price et al., 1991; Snedeker & Trueswell, 2003; Steinhauer et al., 1999). Importantly, this relation is not a one-to-one mapping between individual acoustic cues and syntactic boundaries. Work on English and Mandarin has shown systematic but variable correspondences between syntactic and prosodic phrasing, while perceptual studies show that boundary judgments depend jointly on acoustic cues, syntactic constituency, and prominence (Kuang et al., 2022b;a; 2023). These findings motivate treating acoustic realization, prosodic structure, and linguistic interpretation as related but distinguishable components of the form–function mapping.

Prosodic information in speech representations. Modern speech models increasingly retain information beyond lexical content. SALMONN connects pretrained speech and audio encoders to a language model (Tang et al., 2024); pGSLM jointly models discovered speech units and prosodic feature streams (Kharitonov et al., 2022); and Spirit LM augments speech units with pitch and style information (Nguyen et al., 2025). ProsodyLM explicitly introduces word-level prosody representations and evaluates capabilities including stress understanding and contrastive-focus generation (Qian et al., 2025). Transcription-oriented models can also be adapted to recognize lexical, phrasal, and contrastive stress (Sohn et al., 2025). Together, these studies show that prosodic information can be preserved, represented, or decoded. The present question is stronger: whether such information is grounded in the particular acoustic realization and functionally connected to the linguistic interpretation it is meant to constrain.

Evaluating prosodic structure and communicative function. Recent benchmarks increasingly evaluate prosodic capabilities beyond recognition accuracy or global speech quality. ProsAudit tests structural prosodic knowledge in self-supervised speech models (de Seyssel et al., 2023); EmphAssess evaluates emphasis preservation in speech resynthesis and translation (de Seyssel et al., 2024); StressTest tests interpretations that depend on sentence stress (Yosha et al., 2026); and MMSU includes prosodic phenomena within a broader spoken-language understanding benchmark (Wang et al., 2026). Complementary work on speech generation has shown that TTS systems differ substantially in their realization of prosodic functions and that naturalness or acoustic similarity need not predict successful transmission of the intended meaning (**?**Kuang et al., 2026; Chan & Kuang, 2025). This motivates a function-based evaluation in which generated prosody is assessed not only by acoustic differentiation or perceptual quality, but by whether listeners recover the intended communicative interpretation.

From information presence to functional use. The distinction between information being available and being behaviorally used is central to representation analysis. Probe performance can establish that information is recoverable without establishing that it contributes to the model's own decision process (Hewitt & Liang, 2019). Likewise, model behavior can be sensitive to task presentation and response-format regularities (Sclar et al., 2024). We therefore combine complementary diagnostics rather than relying on a single accuracy measure. Matched productions test whether changing prosody while holding lexical content constant changes interpretation; answer-order reversal separates sensitivity to meaning from response-position preferences; measured acoustics test whether reported prosodic analyses are grounded in the signal; and target-blind human judgments test whether acoustic differences in generated speech successfully transmit the intended meaning. Model explanations are treated as observable outputs whose grounding can be evaluated, not as direct readouts of internal computation.

3 EXPERIMENTAL DESIGN

3.1 STIMULI

We used five syntactically ambiguous English sentences. In each case, the same word sequence supports two distinct syntactic parses, and prosodic phrasing can favor one interpretation over the other. A stronger prosodic boundary at the point where the parses diverge favors one structural analysis, whereas more integrated phrasing favors the competing analysis.

Table 1: Syntactic-ambiguity stimuli and intended interpretations.

Sentence	**Meaning A**	**Meaning B**
The servant of the actress who was crying left.	The actress was crying.	The servant was crying.
Old men and women gathered in the hall.	Both men and women were old.	Only the men were old.
Put the knife on the towel in the box.	The towel was in the box.	Put the knife into the box.
David cleaned the mirror in the hallway.	The mirror was in the hallway.	The cleaning occurred in the hallway.
Lisa saw the man with the telescope.	Lisa used the telescope.	The man had the telescope.

3.2 PRODUCTION

Speech generation. We evaluated six AI production configurations: Gemini TTS, OpenAI TTS, Hume, Nova Sonic, Grok TTS, and Grok STS. Each system generated both intended interpretations of all five sentences, with five independent generation takes per sentence–meaning condition. The intended interpretation was specified using the conditioning procedure supported by each system while keeping the spoken lexical content constant. As a human reference, six native speakers of American English produced both interpretations of all five sentences, yielding 60 recordings.

Acoustic measurements. All human and AI recordings were segmented and aligned at the word level. For each sentence, we identified the word immediately preceding the syntactically critical juncture and extracted four acoustic measures associated with prosodic phrasing: pre-boundary word duration, following pause duration, F0 change across the boundary, and intensity change across the boundary. Word durations and internal pauses across the full utterance were also retained to characterize the location of meaning-dependent timing differences.

Human evaluation. Human listeners evaluated the productions without being told the intended meaning. On each trial, they heard one recording, selected which of the two sentence-specific interpretations they perceived, and rated naturalness on a five-point scale. 32 human listeners contributed 1,917 valid evaluation responses.

3.3 PERCEPTION

AI perception. Six S2S models—Gemini 3.1 Flash Lite, Gemini 3.5 Flash, OpenAI, Grok, Hume EVI, and Nova Sonic—were evaluated as AI listeners. Each model heard the 60 human recordings and selected between the two possible interpretations of the corresponding sentence. Recordings were presented independently and evaluated repeatedly. To assess response-position effects, the complete evaluation was conducted with both the original and reversed ordering of the two alternatives.

Human baseline. Human listeners performed the same two-alternative interpretation task on the human recordings. They heard each utterance without being told the intended interpretation and selected the meaning conveyed by the speaker. Because the lexical content was identical across the two intended meanings, disambiguation depended on information carried by the speech signal.

Model reasoning. All six AI models also produced a brief explanation for each interpretation judgment. We retained these explanations to examine the prosodic evidence reported by each model. Rationales were coded for references to pause or boundary, duration, pitch or intonation, stress or emphasis, and rhythm or grouping. We additionally coded whether the model reported a prosodic boundary at the syntactically critical location, allowing the reported analysis to be compared with the measured acoustics and final interpretation.

4 RESULTS

4.1 PRODUCTION

How reliably could the intended interpretations be distinguished acoustically? For human and AI speech, we trained random forests using following pause, pre-boundary word duration, F0 reset, and intensity change. Human models used leave-one-speaker-out cross-validation, whereas AI models held out generation takes. The four cues reliably distinguished the intended meanings in human speech (85.0% held-out accuracy; AUC = .896), with following pause and pre-boundary lengthening contributing most (Figure 1). AI systems differed substantially in acoustic differentiation (Figure 2). Grok TTS and Hume showed the clearest separation of the intended meanings, whereas Gemini TTS, OpenAI TTS, Nova Sonic, and Grok STS showed much weaker differentiation. Cue profiles also differed across systems: pause and pre-boundary lengthening contributed most for Grok TTS, whereas Hume's contrast was driven primarily by pause. Overall Grok and Hume got the right set of acoustic cues.

Does AI speech place prosodic boundaries at the right locations? Producing the "right" acoustic cues somewhere in the utterance is not sufficient if those cues are associated with the wrong linguistic constituent. We therefore evaluated whether the acoustic contrasts produced by AI systems were aligned with the prosodic structure produced by human, following the two-level framework of Chan & Kuang (2025). The framework distinguishes *structural placement* from *phonetic realization*. A system must first place prosodic structure—for example, boundaries or prominence–on the appropriate words or phrases.

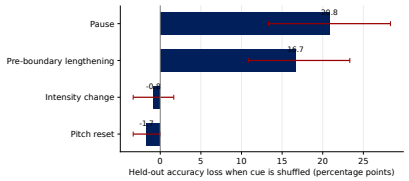

Figure 1: Acoustic cue importance in human production.

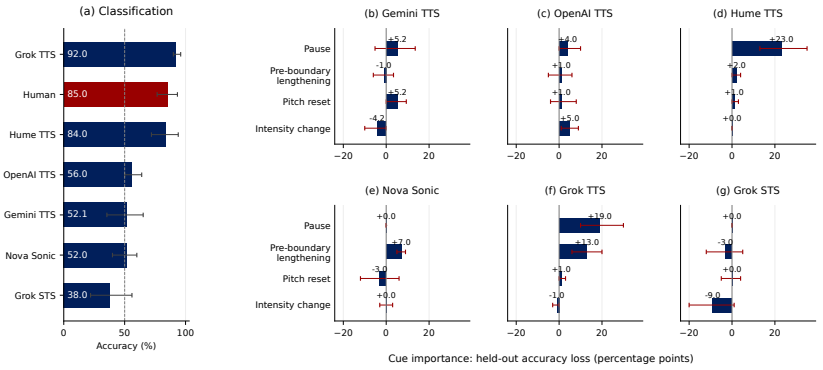

Figure 2: (a) Random-forest classification of intended meanings in human and AI production; (b–e) Random Forest analysis for relative cue weighting in AI production.

For the present phrasing task, we tested structural placement by comparing meaning-dependent pause and word-duration patterns across the entire sentence with the corresponding human contrasts (Table 2). Cosine similarity measures whether the locations of these acoustic changes across lexical positions follow the human pattern. Hume and Grok TTS showed the strongest alignment: Hume for pause ($r_{\cos} = .86$) and Grok TTS for duration ($r_{\cos} = .83$). Gemini showed moderate alignment for pause (.57) but not duration ($-.12$), whereas OpenAI showed modest positive alignment for both (.36 and .22). Nova Sonic and Grok STS showed weak or negative alignment.

Does the generated prosody actually communicate the intended meaning? The strongest test is whether target-blind human listeners recover the intended interpretation. Figure 3 shows that systems with stronger alignment to human production tended also to support better meaning recovery. Grok TTS, which showed strong human alignment in both pause and duration, yielded the highest observed AI meaning-recovery accuracy (72.7%); Hume, also strongly aligned with human timing patterns, reached 60.4%. By contrast, Gemini TTS and OpenAI TTS showed weaker human

Table 2: Whole-sentence alignment of meaning-dependent acoustic changes with human productions. Cosine scores range from -1 (opposite direction) to $+1$ (same direction). The knife sentence was excluded because of an interpretation mismatch in the production prompts.

Production system	Pause	Duration
Gemini TTS	0.57	−0.12
OpenAI TTS	0.36	0.22
Hume	**0.86**	0.79
Nova Sonic	−0.02	−0.40
Grok TTS	0.74	**0.83**
Grok STS	−0.18	−0.003

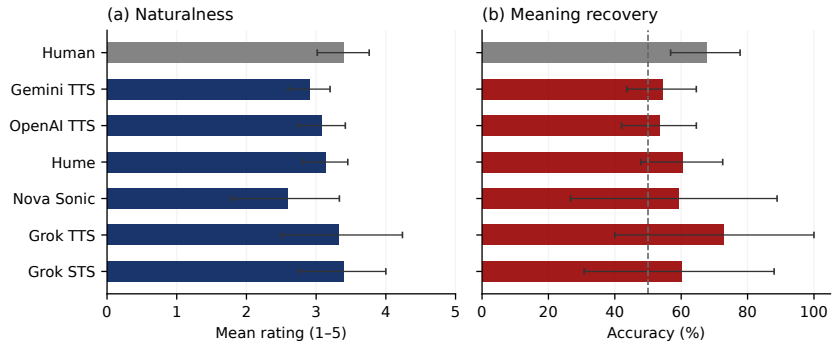

Figure 3: Human evaluation of ambiguity productions from AI systems: (a) mean naturalness rating and (b) intended-meaning recovery accuracy. Error bars indicate 95% crossed listener–recording bootstrap confidence intervals; the dashed line marks 50% chance accuracy.

alignment and remained close to chance (54.3% and 53.4%). This correspondence was not exact, but suggests that human-consistent prosodic realization is functionally relevant to communicative success. Naturalness, however, captured a different property. Grok STS received the highest mean naturalness rating among the AI systems (3.40/5) but only 60.0% meaning recovery, whereas Grok TTS achieved substantially higher recovery (72.7%) with similar naturalness (3.32). Thus, natural-sounding speech is not necessarily communicatively accurate.

Overall, successful prosodic production requires mapping intended meaning onto an appropriate prosodic organization, realizing that organization through suitable acoustic cues at the relevant linguistic locations, and ultimately producing speech from which human listeners can recover the intended meaning.

4.2 Perception

How well can AI models recover ambiguous meanings from prosody? Human listeners recovered the intended interpretation with 90.0% accuracy and macro-F1 of .900, with nearly symmetric performance across the two meanings. Speech models were substantially less reliable (Figure 4); Gemini 3.5 performed best, while the remaining systems showed lower balanced performance and strong interpretation preferences. Overall accuracy can therefore overestimate prosodic understanding: a model may obtain correct answers simply by favoring one interpretation. We instead measured *matched prosodic discrimination*, asking whether changing prosody while holding the lexical sequence constant changed the model's interpretation:

$$D = P(\text{choose A} \mid A\text{-intended prosody}) - P(\text{choose A} \mid B\text{-intended prosody}).$$

Here, $D = 0$ indicates no systematic discrimination between the two prosodic realizations despite any overall response preference. Gemini 3.5 showed the clearest discrimination ($D = .310$), OpenAI a weaker effect ($D = .117$), and Gemini 3.1 only a small net effect ($D = .047$); Nova Sonic, Hume EVI, and Grok showed little stable discrimination. Reversing answer order further showed that these biases had different sources: Hume EVI nearly reversed with option position, Gemini 3.5

Table 3: **Alignment from acoustic evidence to reported prosodic structure and interpretation.** Values are within-sentence Pearson correlations.

Model	Pause ↔ report	Target ↔ report	Report ↔ choice
OpenAI	**.725**	**.466**	**.731**
Gemini 3.5 Flash	.202	.403	**.891**
Gemini 3.1 Flash Lite	.276	.207	.315
Grok	.025	.034	.380
Hume EVI	-.258	-.269	.119
Nova Sonic	-.179	-.078	-.182

and Grok shifted substantially, OpenAI and Nova Sonic changed only modestly, and Gemini 3.1 remained largely stable. Thus, only some models systematically changed their interpretation when prosody changed.

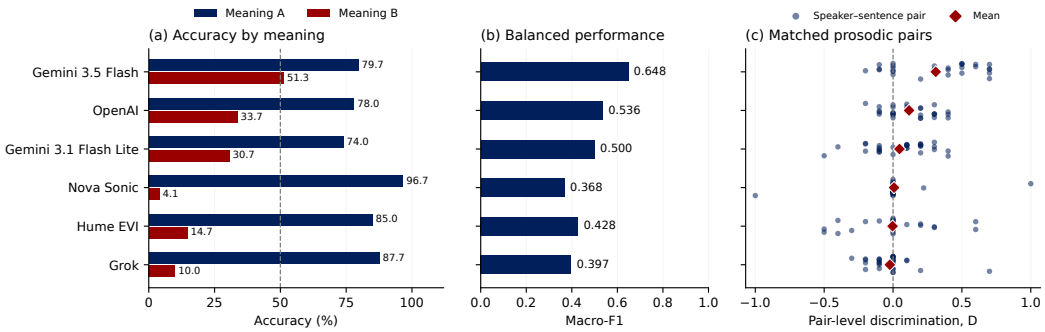

Figure 4: AI interpretation of human ambiguity productions. (a) Accuracy by intended meaning; the dashed line marks 50% chance. (b) Macro-F1 across meanings. (c) Discrimination D for each of 30 matched speaker–sentence pairs; red diamonds indicate model means.

Which acoustic cues predict human interpretation? To establish a human perceptual reference, we trained random forests to predict listeners' interpretations from following pause, pre-boundary duration, F0 reset, and intensity change, holding out both listeners and stimulus groups. Prediction accuracy was 61.2% for human speech and 63.6% for AI speech. Following pause had the largest estimated contribution in both cases, consistent with production (c.f. Figure 1.

Are model explanations grounded in acoustic evidence? To identify where the form-to-function mapping breaks down, we compared model-generated rationales with the measured acoustics and final interpretations of the same recordings. Rationales are treated as observable outputs rather than direct readouts of internal computation.

Models often verbalized linguistically relevant prosodic information. Gemini 3.5, Gemini 3.1, OpenAI, and Grok frequently referred to boundaries or grouping, while Nova Sonic frequently mentioned both boundary and stress information (Figure 5). Thus, weak behavioral performance generally cannot be explained by an inability to state which prosodic properties may be relevant.

We next tested whether these reported analyses were actually grounded. Because following pause was the strongest measured predictor of human interpretation, we used it as the acoustic diagnostic. For each recording, repeated runs yielded a *boundary-report rate*, which we related to (i) measured pause, testing *acoustic grounding*; (ii) target structure, testing *prosodic–linguistic alignment*; and (iii) the model's final interpretation, testing *functional mapping*. Variables were centered within sentence before computing recording-level correlations.

The resulting profiles reveal different failure points. OpenAI showed the strongest acoustic grounding: its reported boundaries tracked measured pause and were also related to both target structure and final interpretation. Yet its overall meaning recovery remained modest, showing that recovering an informative acoustic cue is not sufficient for successful interpretation.

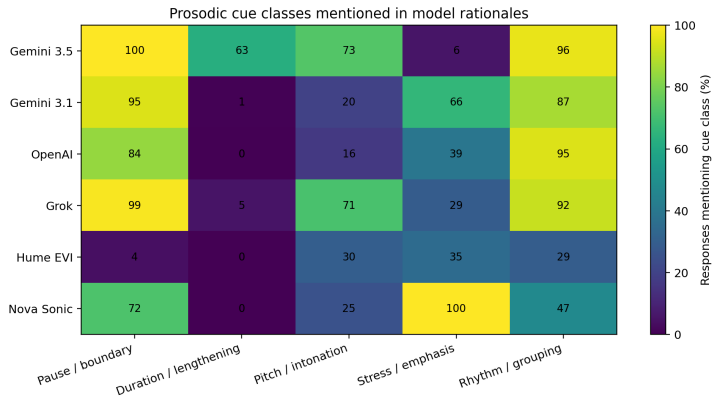

Figure 5: **Prosodic cues mentioned in model rationales.** Cue mention does not imply correct identification or functional use.

Gemini 3.5 showed a different dissociation. Its boundary reports were only weakly related to measured pause, but were more closely aligned with target structure and very strongly coupled to its final choice. Its stronger behavioral discrimination therefore coexisted with weaker grounding of the reported prosodic analysis in the measured signal.

The remaining systems showed weaker alignment across the chain. Gemini 3.1 exhibited only modest coupling; Grok frequently used appropriate prosodic terminology despite almost no acoustic grounding; and Hume EVI and Nova Sonic showed little evidence that reported boundary structure systematically tracked either the signal or the intended interpretation.

Overall, perception requires more than access to acoustic information: the signal must be mapped onto an appropriate prosodic representation, and that representation must in turn constrain linguistic interpretation. The models dissociate at both mappings, showing that prosodic knowledge, acoustic grounding, and functional use are related but distinct competencies.

5 Discussion

The central finding is that prosodic competence does not behave as a single capability. Across models, we observe dissociations among sensitivity to acoustic variation, grounding of prosodic analyses in the signal, mapping between prosodic and linguistic structure, and successful recovery of communicative meaning. These distinctions are analytical rather than a claim that either humans or models implement a fixed sequence of discrete processing stages. Their importance is functional: information may be present in the speech signal, detectable by a model, or even correctly described in a generated explanation without being used reliably to recover meaning.

The perception results make these dissociations particularly clear. Matched prosodic pairs show that overall accuracy can conceal stable interpretation preferences: only some models systematically changed their judgments when prosody changed while lexical content remained fixed. The grounding analysis further shows that successful behavior can break down at different mappings. A model may track an informative acoustic cue yet fail to translate that cue reliably into structural interpretation, or it may produce a prosodic analysis that strongly predicts its own answer while being only weakly related to the measured signal. Conversely, models can verbalize linguistically appropriate prosodic knowledge while their judgments remain largely ungrounded. Thus, acoustic sensitivity, verbalized knowledge, grounding, and functional use provide different kinds of evidence about representation and should not be treated as interchangeable. Table 4 summarizes these distinct model profiles and shows that similar overall performance can arise from different combinations of strengths and failures across these mappings.

Production reveals the complementary direction of the same problem. Generating different acoustic patterns for different intended meanings is not itself sufficient: the relevant acoustic cues must be appropriately organized relative to linguistic structure and ultimately produce an interpretation re-

Table 4: **Representational profiles in prosodic ambiguity perception.** Qualitative summaries integrate behavioral discrimination, acoustic-cue sensitivity, rationale analyses. "Knowledge" refers to prosodic relationships verbalized in model rationales, not direct access to internal representations.

Model	Prosodic discrimination	Verbalized prosodic knowledge	Acoustic grounding	Report–choice coupling	Primary diagnostic limitation
Gemini 3.5	Strongest	Strong	Partial	Strong	Genuine prosodic discrimination, but reported acoustic analysis is not consistently grounded in the measured signal.
OpenAI	Moderate	Moderate–strong	Strongest for following pause	Moderate–strong	Tracks an important human cue reliably, but cue grounding does not consistently yield the intended structural interpretation.
Gemini 3.1	Weak	Strong	Weak	Weak / unstable	Relevant prosodic knowledge is present, but acoustic evidence has limited influence relative to a stable interpretation preference.
Grok	Very weak	Strong	Very weak	Weak	Rich prosodic terminology is largely disconnected from measured acoustic evidence and behavioral discrimination.
Hume EVI	Very weak	Limited	Limited evidence	Weak	Responses are strongly affected by answer position, with little evidence that the critical prosodic contrast systematically controls interpretation.
Nova Sonic	Very weak	Moderate	Very weak	Very weak	Frequently reports prosodic properties, but those reports show little alignment with measured acoustics, target structure, or final interpretation.

coverable by a listener. In our data, systems whose timing patterns were more aligned with human production tended descriptively to support better meaning recovery, while naturalness followed a partly different pattern. Thus, successful prosodic generation depends not simply on producing variation or natural-sounding speech, but on mapping intended meaning onto an appropriate prosodic organization, realizing that organization acoustically, and transmitting the intended meaning to a listener.

Taken together, production and perception motivate a functional view of prosodic representation:

$$\text{acoustics} \leftrightarrow \text{prosodic organization} \leftrightarrow \text{linguistic structure} \leftrightarrow \text{communicative meaning}.$$

Perception asks whether variable acoustic evidence can be mapped through prosodic and linguistic organization to meaning; production asks whether an intended meaning can be mapped in the opposite direction into an acoustic realization from which a listener can recover it. The key criterion is therefore not whether acoustics, prosody, or linguistic structure are represented individually, but whether these representations are *functionally aligned*. Human listeners provide a useful functional reference precisely because they recover relatively stable interpretations despite substantial variability in acoustic realization. Current speech foundation models show partial competence at different links in this mapping, but those links are not yet reliably coordinated.

AI USE STATEMENT

In this work, we used generative AI tools to assist with identifying literature gap, support qualitative analysis of model-generated rationales, modifying figures, clean and format datasets, editing and proofreading drafts for clarity, formatting references and tables.

We did not use generative AI tools to generate datasets, conceptualize the work, design experiment, proposing or refining hypotheses, generate synthetic datasets, formulate or prove mathematical claims, perform translation.

All AI-assisted work was reviewed by the authors. Suggestions concerning analyses and interpretation were checked against the underlying experimental data and statistical results, and AI-suggested literature and citations were verified against the original sources. The authors made all final decisions regarding experimental design, analyses, scientific claims, and interpretation.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study includes human speech production and perception experiments. Human-subject procedures were conducted under institutional ethical oversight, with informed consent and appropriate protections for participant data. The experiments involve controlled linguistic judgments and do not collect sensitive personal information beyond that required for the speech study. Model outputs are analyzed for research purposes only. Any released human data will follow the consent and sharing conditions of the approved protocol, with identifying information removed where applicable. Generated speech and model responses will be released only where permitted by the relevant model providers' terms and licenses.

REPRODUCIBILITY STATEMENT

We document the complete stimulus set, intended interpretations, model prompts, model identifiers and access dates, generation and evaluation procedures, acoustic feature definitions, rationale-coding procedures, and statistical analyses in the paper. We will release analysis code, trial-level processing procedures, and derived measurements needed to reproduce the reported figures, tables, and statistical results, subject to participant consent and model-provider terms. Because several evaluated systems are proprietary hosted models, exact reproduction may depend on continued access to the corresponding model versions; we therefore report model versions, settings, and access conditions as precisely as possible.

REFERENCES

Cedric Chan and Jianjing Kuang. Exploring the accuracy of prosodic encodings in state-of-the-art text-to-speech models. In *Proceedings of the 12th International Conference on Speech Prosody (Speech Prosody 2024)*, pp. 27–31, 2024. doi: 10.21437/SpeechProsody.2024-6. URL https://doi.org/10.21437/SpeechProsody.2024-6.

Cedric Chan and Jianjing Kuang. Toward objective and interpretable prosody evaluation in text-to-speech: A linguistically motivated approach. *arXiv preprint arXiv:2511.02104*, 2025.

Maureen de Seyssel, Marvin Lavechin, Hadrien Titeux, Arthur Thomas, Gwendal Virlet, Andrea Santos Revilla, Guillaume Wisniewski, Bogdan Ludusan, and Emmanuel Dupoux. ProsAudit, a prosodic benchmark for self-supervised speech models. In *Interspeech 2023*, pp. 2963–2967, 2023. doi: 10.21437/Interspeech.2023-438. URL https://www.isca-archive.org/interspeech_2023/deseyssel23_interspeech.html.

Maureen de Seyssel, Antony D'Avirro, Adina Williams, and Emmanuel Dupoux. EmphAssess: a prosodic benchmark on assessing emphasis transfer in speech-to-speech models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 495–507, 2024. doi: 10.18653/v1/2024.emnlp-main.30. URL https://aclanthology.org/2024.emnlp-main.30/.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Morgane Riviere, Abdelrahman Mohamed, Emmanuel Dupoux, and Wei-Ning Hsu. Text-free prosody-aware generative spoken language modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8666–8681, 2022. doi: 10.18653/v1/2022.acl-long.593. URL https://aclanthology.org/2022.acl-long.593/.

Jianjing Kuang, May Pik Yu Chan, and Nari Rhee. The effects of syntactic and acoustic cues on the perception of prosodic boundaries. In *Proceedings of Speech Prosody*, pp. 699–703, 2022a.

Jianjing Kuang, May Pik Yu Chan, Nari Rhee, Mark Liberman, and Hongwei Ding. The mapping between syntactic and prosodic phrasing in English and Mandarin. In *Proceedings of Interspeech 2022*, pp. 3443–3447, 2022b. doi: 10.21437/Interspeech.2022-10657. URL https://doi.org/10.21437/Interspeech.2022-10657.

Jianjing Kuang, May Pik Yu Chan, and Nari Rhee. Boundary perception as interplay by acoustic cues, syntactic constituency, and prominence. In *Proceedings of the 20th International Congress of Phonetic Sciences*, 2023.

Jianjing Kuang, Ethan Yang, and Kevin Li. Evaluating prosodic encodings of information structure in generative speech AI. In *Proc. Speech Prosody 2026*, pp. 144–148, Philadelphia, PA, USA, 2026. doi: 10.21437/SpeechProsody.2026-29.

D. Robert Ladd. *Intonational Phonology*. Cambridge University Press, 2 edition, 2008.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoît Sagot, and Emmanuel Dupoux. SpiRit-LM: Interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13:30–52, 2025. doi: 10.1162/tacl_a_00728. URL https://aclanthology.org/2025.tacl-1.2/.

Janet B. Pierrehumbert and Julia Hirschberg. The meaning of intonational contours in the interpretation of discourse. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack (eds.), *Intentions in Communication*, pp. 271–311. MIT Press, 1990.

Patti Price, Mari Ostendorf, Stefanie Shattuck-Hufnagel, and Cynthia Fong. The use of prosody in syntactic disambiguation. In *Proceedings of the Workshop on Speech and Natural Language*, pp. 372–377, 1991.

Kaizhi Qian, Xulin Fan, Junrui Ni, Slava Shechtman, Mark A. Hasegawa-Johnson, Chuang Gan, and Yang Zhang. ProsodyLM: Uncovering the emerging prosody processing capabilities in speech language models. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=uBg8PClMUu.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6c0e99d736da621403018ca7b32b1a4d-Abstract-Conference.html.

Jesse Snedeker and John C. Trueswell. Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, 48(1):103–130, 2003.

Samuel S. Sohn, Sten Knutsen, and Karin Stromswold. Harnessing Whisper for prosodic stress analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25931–25942, 2025. doi: 10.18653/v1/2025.findings-acl.1331. URL https://aclanthology.org/2025.findings-acl.1331/.

Karsten Steinhauer, Kai Alter, and Angela D. Friederici. Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience*, 2(2):191–196, 1999. doi: 10.1038/5757. URL https://doi.org/10.1038/5757.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/476ab8f369e489c04187ba84f68cfa68-Abstract-Conference.html.

Michael Wagner and Duane G. Watson. Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, 25(7–9):905–945, 2010.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. MMSU: A massive multi-task spoken language understanding and reasoning benchmark. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/34e278fbbd7d6d7d788c98065988e1a9-Abstract-Conference.html.

Iddo Yosha, Gallil Maimon, and Yossi Adi. StressTest: Can YOUR speech LM handle the stress? In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1264–1285, 2026. doi: 10.18653/v1/2026.findings-acl.64. URL https://aclanthology.org/2026.findings-acl.64/.