Modeling Sarcastic Speech: Semantic and Prosodic Cues in a Speech Synthesis Framework

Zhu Li[1], **Yuqing Zhang**[2], **Xiyuan Gao**[1], **Shekhar Nayak**[1] **& Matt Coler**[1]
[1]Speech Technology Lab, University of Groningen, Campus Fryslân, The Netherlands
[2]Center for Language and Cognition, University of Groningen, The Netherlands

Abstract

Sarcasm is a pragmatic phenomenon in which speakers convey meanings that diverge from literal content, relying on an interaction between semantics and prosodic expression. However, how these cues jointly contribute to the recognition of sarcasm remains poorly understood. We propose a computational framework that models sarcasm as the integration of semantic interpretation and prosodic realization. Semantic cues are derived from an LLaMA 3 model fine-tuned to capture discourse-level markers of sarcastic intent, while prosodic cues are extracted through semantically aligned utterances drawn from a database of sarcastic speech, providing prosodic exemplars of sarcastic delivery. Using a speech synthesis testbed, perceptual evaluations show that semantic and prosodic cues enhance perceived sarcasm, with the combined system achieving the best downstream F1 while maintaining high subjective sarcasm ratings. These findings highlight the complementary roles of semantics and prosody in pragmatic interpretation and illustrate how modeling can shed light on the mechanisms underlying sarcastic communication.

Keywords: pragmatics; sarcasm; prosody; semantic cues; speech perception

Introduction

Sarcasm plays an important role in everyday communication by enabling speakers to express mockery, criticism, or contempt through meanings that diverge from literal interpretation. As a pragmatic phenomenon, sarcasm is often realized through irony, where ostensibly positive or neutral expressions convey an underlying negative intent. While such ironic intent can sometimes be inferred from explicit linguistic content, sarcasm frequently lacks clear lexical markers and therefore cannot be reliably identified from semantics alone. Instead, its interpretation depends on a nuanced interplay of semantic incongruity, contextual information, and prosodic cues. Speakers commonly employ prosodic strategies, such as exaggerated intonation, emphatic stress, or lengthened syllables, to signal sarcastic intent, particularly in cases where linguistic cues are subtle or ambiguous (Cheang & Pell, 2008; Z. Li et al., 2024; Rockwell, 2000). This reliance on multiple cues makes sarcasm a complex phenomenon that challenges computational modeling, as successful automatic recognition requires integrating information across modalities. Understanding how listeners integrate these cues is a fundamental question for theories of pragmatic language comprehension.

One challenge in studying sarcastic communication lies in systematically manipulating semantic and prosodic cues while holding other factors constant. Traditional corpus-based or behavioral studies often lack fine-grained control over semantic and prosodic cues, since these dimensions tend to be entangled in natural speech and are difficult to manipulate independently or in a graded manner as experimental stimuli (Bryant, 2010; Bryant & Fox Tree, 2002; Rockwell, 2007). Speech synthesis, however, offers a powerful experimental testbed, enabling controlled generation of utterances in which semantic and prosodic information can be independently varied to explore their respective contributions to sarcasm perception.

Recent advances in computational modeling provide new opportunities to formally represent and manipulate the cues underlying pragmatic communication. Large language models (LLMs) have shown an ability to capture discourse-level semantic and pragmatic patterns, including forms of semantic subtlety relevant to sarcasm. At the same time, retrieval-based approaches allow models to condition generation on prosodic exemplars that reflect expressive sarcasm patterns. Together, these methods offer a way to formalize hypotheses about how semantic content and prosodic realization jointly contribute to sarcastic communication.

In this study, we introduce a sarcasm-aware speech synthesis framework that models sarcasm as the integration of semantic content and prosodic expression. We independently manipulate two types of cues to condition sarcastic speech generation in our synthesis framework: 1) semantic representations of sarcastic intent derived from a fine-tuned LLM, and 2) prosodic patterns drawn from aligned reference utterances in a curated speech database of sarcastic speech. Using perceptual evaluations of the synthesized speech, we assess how each cue contributes to listeners' perception of sarcasm. Perceptual evaluations reveal that both cues independently enhance sarcasm recognition, with the strongest effects emerging when they are combined. These findings provide computational evidence for the complementary roles of semantics and prosody in pragmatic interpretation and demonstrate how synthesis-based models can serve as tools for modeling speech behavior.

Related Work

Sarcasm Detection and Modeling

In recent years, substantial progress has been made in computational sarcasm detection. Many studies have adopted multimodal approaches that integrate textual, acoustic, and visual cues, demonstrating significant improvements over unimodal

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

methods (Castro et al., 2019; Gao et al., 2024; Z. Li, Gao, Zhang, et al., 2025; Raghuvanshi et al., 2025; Ray et al., 2022). Datasets such as MUStARD (Castro et al., 2019) and its extensions (Ray et al., 2022) have facilitated this line of work by providing aligned textual, auditory, and visual data for sarcasm recognition. While these efforts highlight the feasibility of modeling sarcasm computationally, they have largely focused on classification and recognition tasks. Recent work has quantified how pragmatic information is distributed across text and prosody. For example, Yadavalli et al. (2025) show that prosody carries substantial information about sarcasm when long-term discourse context is unavailable. However, relatively little attention has been paid to the generation of sarcastic speech or to using computational models as tools for exploring how different cues contribute to sarcasm perception. Moving from detection to synthesis is not only a natural next step but also a crucial one, as it enables interactive systems to actively deploy pragmatic phenomena rather than passively identify them.

Expressive Speech Synthesis and Sarcasm

Research on expressive text-to-speech (TTS) synthesis has primarily concentrated on broad emotional categories such as happiness, anger, and sadness (Akuzawa et al., 2018; T. Li et al., 2021; Wang et al., 2018). Although these systems can produce emotionally expressive and intelligible speech, more subtle and pragmatic phenomena remain largely unexplored. Sarcasm, despite being widely used in everyday communication, has received particularly little attention in speech synthesis research. However, sarcastic or other pragmatic speech synthesis holds considerable potential for improving human-computer interaction in applications such as conversational agents and entertainment systems (Ritschel et al., 2019). Prior work has attempted to analyze acoustic correlates of sarcastic speech, identifying prosodic features related to pitch, pace, and loudness (Cheang & Pell, 2008; Z. Li et al., 2024). However, only a limited number of studies have explored directly manipulating such prosodic features to generate sarcastic-sounding speech (Peters & Almor, 2017). This difficulty arises in part because sarcasm is inherently more subtle and context-dependent than conventional emotions such as anger or joy, making it difficult to capture with handcrafted acoustic features or simple prosodic controls. Moreover, the development of sarcastic speech synthesis systems is hindered by the scarcity of high-quality, annotated sarcastic speech corpora, which limits the applicability of purely data-driven approaches (Z. Li, Zhang, et al., 2025; Z. Li et al., 2023).

LLM-Based Semantics and Retrieval-Based Speech Synthesis

Recent advances in natural language processing have introduced new opportunities for modeling pragmatic phenomena. LLMs have demonstrated the ability to capture high-level semantic and discourse-level information. In the TTS domain, integrating LLM-derived embeddings has been shown to enable more nuanced prosody control and improve emotional expressiveness and contextual appropriateness (Feng & Yoshimoto, 2024; S. Li et al., 2024; Shen et al., 2025). Such semantic representations are particularly relevant for sarcasm, which often depends on pragmatic incongruity between literal meaning and intended attitude. Leveraging LLM-based semantic representations could therefore provide a principled way of conditioning TTS models on sarcasm-relevant cues that handcrafted features fail to capture.

Complementary to semantic modeling, retrieval-based methods have been explored as a way to condition speech generation on external knowledge or examples retrieved from large databases. Retrieval-augmented generation (RAG) has been widely applied in text-based tasks such as question answering and dialogue (Lewis et al., 2020; Shuster et al., 2021), and related ideas have been adopted in speech synthesis through reference-based and style-transfer approaches, where reference utterances are used to guide prosodic realization (Luo et al., 2025; Xue et al., 2024). However, existing approaches often rely on manually selected or speaker-specific reference audio, limiting scalability and contextual alignment. In the context of sarcasm, where annotated corpora are scarce, automatic retrieval of semantically similar sarcastic speech could provide prosodic exemplars that both enrich training and guide generation in data-scarce settings. This suggests a paradigm in which LLMs supply semantic-pragmatic cues while the RAG module retrieves expressive references for prosodic encoding, together enabling a systematic modeling of subtle pragmatic phenomena such as sarcasm.

Methods

We propose a Retrieval-Augmented LLM-enhanced sarcastic speech modeling framework. An overview of the architecture is illustrated in Figure 1.

Sarcasm-Aware Semantic Encoding via Parameter-Efficient Fine-tuning (PEFT)

To systematically manipulate semantic cues of sarcasm, we encode input text into sarcasm-aware semantic representations using a PEFT method on a sarcasm-labeled corpus. This allows us to isolate and represent high-level semantic features indicative of sarcastic intent for subsequent synthesis and perceptual evaluation.

Low-Rank Adaptation (LoRA) has recently been widely adopted for adapting LLMs across diverse domains such as emotion recognition, sentiment analysis, and affective detection, due to its parameter efficiency and effectiveness (Cai et al., 2024). Motivated by these advances, we leverage LoRA to finetune the LLaMA 3-8B model for sarcasm-aware semantic encoding. Fine-tuning is performed on a curated sarcasm-labeled dataset, enabling the model to capture signals of sarcastic intent, including pragmatic and discourse-level cues.

Formally, given an input text sequence $x = (w_1, \ldots, w_{T_t})$ of length T_t, the LLaMA 3-LoRA encoder produces a sequence

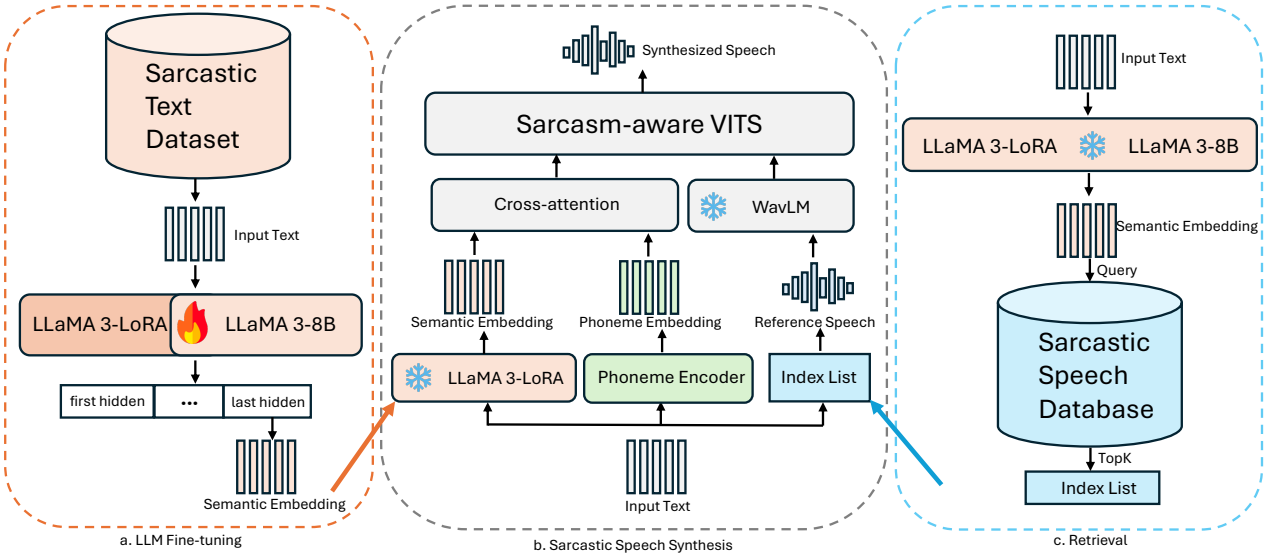

Figure 1: Overview of the proposed framework modeling sarcasm as an interaction between semantic and prosodic cues. Semantic representations capturing semantic and pragmatic cues are extracted from a sarcasm-adapted LLaMA 3 model, while prosodic exemplars are retrieved from reference sarcastic speech. These cues are integrated within a speech synthesis architecture to generate utterances that convey sarcastic intent.

of contextualized semantic embeddings:

$$\mathbf{E}_s = f_{\text{LoRA}}(x) \in \mathbb{R}^{T_t \times d_t}, \tag{1}$$

where d_t denotes the hidden dimensionality of the language model. Each row of $\mathbf{E}_s$ corresponds to a token-level representation extracted from the hidden state of the final transformer layer, encoding meanings relevant to sarcasm and providing a controllable semantic representation for sarcasm synthesis.

RAG for Prosody Conditioning

To investigate the role of prosodic cues in sarcasm perception, we retrieve semantically aligned sarcastic utterances as prosodic exemplars. These exemplars allow us to manipulate prosody independently of semantic content for synthesis.

We integrate an RAG module into the sarcastic speech synthesis framework. The key idea is to leverage semantically aligned sarcastic utterances as prosodic references, providing the generated speech with realistic intonation patterns. Concretely, we first construct an index $\mathcal{D} = \{(u_i, \mathbf{a}_i)\}_{i=1}^{N}$ of sarcastic utterances u_i curated from MUStARD++, where $\mathbf{a}_i$ denotes their pre-computed semantic representations using the same LLaMA 3-LoRA adapted LLaMA 3-8B encoder. For a given input text, the semantic embedding $\mathbf{E}_s$ produced by the LLaMA 3-LoRA encoder is used as a retrieval query to fetch the top-K semantically relevant sarcastic utterances from the indexed database; in this work, we set $K = 3$.

We first compute the cosine similarity between $\mathbf{E}_s$ and each database entry:

$$\text{sim}(\mathbf{E}_s, \mathbf{a}_i) = \frac{\mathbf{E}_s \cdot \mathbf{a}_i}{\|\mathbf{E}_s\| \|\mathbf{a}_i\|}. \tag{2}$$

We then retrieve the top-K most relevant utterances. Each retrieved utterance $u_k \in \mathcal{U}_{\text{top-}K}$ is encoded by WavLM (Chen et al., 2022) to obtain a prosody embedding:

$$\mathbf{E}_{w_k} = \text{Pool}(f_{\text{WavLM}}(u_k)) \in \mathbb{R}^{d_w}, \tag{3}$$

where $\text{Pool}(\cdot)$ aggregates the frame-level features into a fixed-length vector, and d_w denotes the dimensionality of the resulting prosodic embedding. These exemplars provide style references that reflect the characteristic intonation and emphasis patterns of sarcastic speech.

Compared to conventional style transfer methods that rely on a single manually chosen reference utterance, the proposed retrieval mechanism offers two key advantages. First, it scales more naturally to diverse input contexts, since the retrieval process automatically identifies semantically aligned exemplars. Second, by conditioning on multiple retrieved samples, the model can capture a richer variety of sarcastic prosodic patterns, leading to more contextually appropriate speech.

LLM-Enhanced and Retrieval-Guided TTS for Sarcasm

The system builds upon the VITS architecture (Kim et al., 2021), enriched with semantic features extracted by a fine-tuned LLaMA 3 and guided by prosodic exemplars retrieved via a RAG module.

Let $\mathbf{E}_p \in \mathbb{R}^{T_p \times d_p}$ denote the sequence of phoneme embeddings, where T_p is the number of phonemes in the input sequence and d_p is the dimensionality of the phoneme embedding space. Let $\mathbf{E}_s \in \mathbb{R}^{T_t \times d_t}$ denote the sarcasm-aware semantic textual embeddings produced by the fine-tuned LLaMA 3. In the cross-attention module, we treat $\mathbf{E}_p$ as the query and $\mathbf{E}_s$ as key and value:

$$\mathbf{Q} = \mathbf{E}_p \mathbf{W}_q, \quad \mathbf{K} = \mathbf{E}_s \mathbf{W}_k, \quad \mathbf{V} = \mathbf{E}_s \mathbf{W}_v, \tag{4}$$

$$\mathbf{H} = \text{Softmax}\Big(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\Big)\mathbf{V}, \tag{5}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ are learned projection matrices, and d_k is the dimensionality of the key. The resulting cross-attention output $\mathbf{H}$ represents a phoneme-aligned semantic encoding.

Finally, the prosody exemplars $\mathcal{E}_w = \{\mathbf{E}_{w_1}, \ldots, \mathbf{E}_{w_K}\}$ extracted by WavLM are linearly projected and added to the cross-attention output over semantic and phoneme embeddings to modulate the decoder:

$$\mathbf{Z} = \mathbf{H} + \sum_{k=1}^{K} \mathbf{W}_w \mathbf{E}_{w_k}, \tag{6}$$

where $\mathbf{W}_w$ is a learned linear projection mapping the prosody embeddings to the decoder dimension. This yields hidden states $\mathbf{Z}$ that are conditioned on phonemes, semantic content, and retrieved prosodic cues, enabling the synthesis of sarcastic speech with expressive intonation.

Experiments

To investigate how semantic and prosodic cues contribute to the perception of sarcasm, we conducted a series of experiments using synthesized speech with systematically manipulated cues. We define a set of experimental conditions that independently and jointly vary semantic and prosodic information. Specifically, we examine:

- **Baseline**: speech generated without sarcasm-specific semantic or prosodic guidance;
- **Semantic-only**: speech conditioned on sarcasm-aware semantic embeddings derived from a fine-tuned LLaMA 3 model;
- **Prosody-only**: speech guided by prosodic exemplars retrieved from a database of sarcastic utterances;
- **Combined**: speech incorporating both semantic and prosodic guidance.

In later analyses, we further explore sub-variants within the semantic-only condition, including embeddings from BERT, pretrained LLaMA 3, and LoRA-fine-tuned LLaMA 3, in order to assess the effect of different semantic representations. These conditions collectively allow us to evaluate the relative contributions of semantic cues, prosodic expression, and their interaction to listeners' perception of sarcastic intent.

Data

For the pre-training phase of VITS, we use the HiFi-TTS corpus (Bakhturina et al., 2021), a high-fidelity audiobook dataset with diverse speakers and wideband recordings.

To build the sarcastic speech database for retrieval purposes, we use the MUStARD++ dataset (Ray et al., 2022). MUStARD++ is a multimodal sarcasm detection corpus compiled from popular sitcoms such as Friends and The Big Bang Theory. It contains 1,202 audiovisual utterances, equally divided into 601 sarcastic and 601 non-sarcastic samples. Following standard practice, we partition the corpus into training, validation, and test sets with an 8:1:1 ratio. Specifically, 10% of the data is held out for evaluation, while the remaining 80% of sarcastic samples are used to construct the retrieval-based sarcastic speech database. This setup ensures relatively sufficient coverage of sarcastic prosody in the retrieval dataset.

To adapt LLaMA 3 as sarcasm-aware semantic encoders via LoRA, we fine-tune them on the News Headlines Sarcasm dataset (Misra & Arora, 2023). This dataset contains over 28,000 headlines from sources such as *The Onion*, each annotated with a binary sarcasm label. The class distribution is roughly balanced, making it suitable for learning discourse-level markers of sarcastic intent. The LLM serves as a feature extractor that provides high-level semantic embeddings conditioned on sarcastic intent. To validate the effectiveness of this adaptation, we train a sarcasm detection model using various textual embeddings, including BERT, LLaMA 3, LLaMA 3-LoRA, and compare their performance as evidence of each embedding's ability to capture sarcastic intent.

Experimental Setup

For training the baseline VITS model, we follow the standard configuration described in Kim et al. (2021). The model is pre-trained on the HiFi-TTS corpus (Bakhturina et al., 2021) using the open-source Amphion toolkit[1].

To adapt LLMs for sarcasm-aware synthesis, the LoRA fine-tuning of LLaMA 3 is performed with an expansion factor of 8 and a learning rate of $1e-4$, updating only the low-rank adapters in the attention layers while freezing the backbone weights[2].

Compared Methods

To systematically assess the contribution of semantic and prosodic cues to sarcasm perception, we synthesized speech under six conditions corresponding to different manipulations of these cues:

- **Baseline**: Speech generated using the standard variational text-to-speech model VITS, without any sarcasm-specific semantic or prosodic guidance, serving as a neutral reference.
- **Semantic-only (BERT)**: Speech conditioned on semantic embeddings from a pretrained BERT model (Devlin et al., 2019), providing general contextual semantic features.
- **Semantic-only (LLaMA 3)**: Speech conditioned on embeddings from LLaMA 3 (Dubey et al., 2024), enabling richer contextual and pragmatic information.
- **Semantic-only (LLaMA 3-LoRA)**: Speech conditioned on sarcasm-aware semantic embeddings from a LoRA-fine-tuned LLaMA 3 model, capturing discourse-level cues indicative of sarcastic intent.

[1] https://github.com/open-mmlab/Amphion
[2] https://github.com/hiyouga/LLaMA-Factory

- **Prosody-only (RAG)**: Speech guided by prosodic exemplars retrieved from a database of sarcastic utterances, providing realistic intonation patterns while keeping semantic content neutral.
- **Semantic + Prosody (Proposed)**: Speech generated using both sarcasm-aware semantic embeddings (LLaMA 3-LoRA) and retrieved prosodic exemplars (RAG), allowing joint manipulation of semantic incongruity and prosodic expression to maximize perceived sarcasm.

Results and Discussion

Effect of Different Semantic Embeddings on Sarcasm Intent Modeling

We first evaluate how well different semantic embeddings capture semantic cues associated with sarcastic intent. Specifically, we train and evaluate sarcasm detection models on the text transcripts of the MUStARD++ dataset, using the same 8:1:1 data split and following the same experimental setup as Ray et al. (2022). The goal of this experiment is to assess how effectively different semantic embeddings capture sarcasm-relevant cues and to motivate the choice of embeddings used later for conditioning speech synthesis. Models are evaluated with Precision (P), Recall (R), and weighted F1-score (F1), with results summarized in Table 1.

Table 1: Performance of different models on sarcasm detection. MUStARD++ (text) refers to the original sarcasm detection systems used in Ray et al. (2022).

Method	P (%)	R (%)	F1 (%)
MUStARD++ (text)	67.9	67.7	67.7
BERT	66.8	66.9	66.8
LLaMA 3	65.7	65.4	65.5
LLaMA 3-LoRA	**72.4**	**72.7**	**72.5**

Both the MUStARD++ baseline and BERT achieved moderate performance with F1-scores around 67%. LLaMA 3 performed slightly worse (65.5%), suggesting that general-purpose pretraining alone is insufficient to capture the nuanced cues of sarcasm. In contrast, PEFT with LoRA yielded a clear improvement, raising the F1-score to 72.5%. This indicates that targeted fine-tuning effectively operationalizes sarcasm-relevant semantic features. These results provide evidence that semantic embeddings from the LoRA-enhanced model more reliably encode discourse-level markers of sarcasm. This motivates their use in the TTS framework: by conditioning speech synthesis on embeddings that better reflect sarcastic intent, we can systematically examine whether improvements in semantic cue representation translate to perceptually stronger expressions of sarcasm in synthesized speech.

Sarcastic Speech Synthesis Results

Table 2 summarizes the evaluation of synthesized speech under different cue manipulations: semantic embeddings, prosodic exemplars, and their combination. We consider three perspectives: 1) low-level acoustic consistency, measured by mel-cepstral distortion (MCD), pitch root mean square error (RMSE), energy RMSE; 2) operationalized sarcasm expressivity, evaluated via a downstream sarcasm detection model applied to synthesized audio, and 3) subjective perception of sarcasm, assessed through two mean opinion score (MOS) tests: naturalness MOS (NMOS) and sarcasm MOS (SMOS). To assess downstream sarcasm detection performance, we adopt the detection architecture used in MUStARD++ with collaborative gating (Ray et al., 2022), using only the *speech modality*. Features extracted from synthesized audio are compared against target sarcasm labels, where higher agreement indicates stronger sarcasm expressiveness.

Overall, speech generated with both sarcasm-aware semantic embeddings (LoRA-fine-tuned LLaMA 3) and retrieved prosodic exemplars exhibited the strongest performance across measures. Specifically, the combined condition achieved a comparably lower distortion rate (9.83 MCD) and the most stable prosodic statistics, while also yielding the highest downstream sarcasm detection F1-score (62.5%). Semantic embeddings from the untuned LLaMA 3 alone resulted in higher distortion and lower detection performance, suggesting that general-purpose embeddings fail to fully capture the semantic cues critical for conveying sarcasm. Fine-tuning with LoRA improves semantic alignment, and further incorporating prosodic exemplars enhances the perceptual expressivity of the synthesized utterances. Importantly, low-level acoustic metrics (MCD, pitch, energy) remained relatively stable across conditions, indicating that improvements in sarcasm expressivity are primarily due to high-level semantic and prosodic cue manipulation, rather than changes in basic signal fidelity. Moreover, the increases observed in sarcasm detection closely mirror the subjective ratings, suggesting that the objective recognizability of sarcastic intent reliably translates to perceptual expressivity, as evaluated below.

To complement the measures above, we conducted a perceptual study to evaluate how well the synthesized speech conveyed sarcasm and naturalness. We recruited 30 listeners from diverse backgrounds[3] who were asked to rate randomly shuffled samples from each system on two dimensions: (i) perceived naturalness (NMOS) and (ii) perceived sarcasm expressivity (SMOS), both on 5-point Likert scales. Each participant listened to eight sentences generated by six systems, plus the ground truth, resulting in a total of 56 utterances[4]. Pairwise comparison across conditions reveals several important trends. Baseline speech without semantic or prosodic guidance was generally intelligible but relatively flat, receiving moderate naturalness (2.6 ± 0.1) and sarcasm expressivity (3.2 ± 0.2). Adding general-purpose semantic embeddings (BERT) did not meaningfully improve ratings (2.5 ±

[3]We did not collect demographic variables beyond English proficiency and hearing status; we acknowledge this as a limitation.

[4]Inter-rater reliability for naturalness ratings was modest at the individual-rater level ($ICC_{(2,30)}$ = .86 for NMOS and ($ICC_{(2,30)}$ = .80 for SMOS).

Table 2: Evaluation of sarcastic speech synthesis systems. Objective acoustic metrics (MCD, pitch, energy) assess signal fidelity; sarcasm detection (Precision (P), Recall (R), and weighted F1-score (F1)) evaluates how well sarcastic intent can be recovered from synthesized speech; subjective ratings measure perceived naturalness (NMOS) and sarcasm expressivity (SMOS).

	Cue Type		Objective			Sarcasm Detection			Subjective	
Method	**Semantic**	**Prosodic**	**MCD ↓**	**Pitch ↓**	**Energy ↓**	**P ↑**	**R ↑**	**F1 ↑**	**NMOS ↑**	**SMOS ↑**
Ground Truth	–	–	–	–	–	62.8	62.3	62.3	3.8 ± 0.1	4.5 ± 0.2
Baseline	–	–	9.8	261.9	4.4	60.3	60.5	59.9	2.6 ± 0.1	3.2 ± 0.2
w/ BERT	✓	–	**9.6**	265.6	4.5	60.5	60.6	60.5	2.5 ± 0.1	3.1 ± 0.2
w/ LLaMA 3	✓	–	10.4	261.6	4.4	59.6	59.7	59.0	2.0 ± 0.1	2.6 ± 0.2
w/ LLaMA 3-LoRA	✓	–	10.1	282.6	4.4	60.6	60.8	60.6	2.6 ± 0.1	3.8 ± 0.2
w/ RAG	–	✓	10.0	261.0	4.4	61.5	61.7	61.6	2.6 ± 0.1	3.7 ± 0.2
w/ LoRA + RAG	✓	✓	9.8	**259.6**	4.4	**62.7**	**62.9**	**62.5**	**2.7 ± 0.1**	**3.8 ± 0.2**

0.1 NMOS, 3.1 ± 0.2 SMOS), suggesting that embeddings lacking sarcasm-specific information may be insufficient to strengthen perceptual sarcasm cues. In addition, directly integrating LLaMA 3 degraded performance (2.0 ± 0.1 NMOS, 2.6 ± 0.2 SMOS). Listeners frequently reported that these samples sounded less natural and had flat intonation patterns. This indicates that raw LLM embeddings are poorly aligned with intended expressive speech. In contrast, semantic embeddings adapted for sarcastic content were associated with higher perceived sarcasm ratings (3.8 ± 0.2 SMOS) while maintaining naturalness comparable to the baseline (2.6 ± 0.1 NMOS). Listeners consistently reported clearer prosodic patterns, such as exaggerated intonation, that aligned with the intended sarcastic meaning, suggesting that targeted semantic conditioning via LoRA adaptation can operationalize semantically and pragmatically relevant sarcastic cues in speech. Incorporating prosodic exemplars retrieved from real sarcastic utterances also yielded relatively high sarcasm expressivity ratings (SMOS 3.7–3.8), without reducing naturalness (2.6 ± 0.1 NMOS). Samples were described as having intonation and emphasis patterns more consistent with authentic sarcastic speech, making the intended sarcasm more immediately recognizable by a large portion of participants. Overall, the combined manipulation of semantic and prosodic cues achieved the highest downstream sarcasm-detection F1 and the highest numerical NMOS (2.7 ± 0.1 NMOS) among synthesized systems, while maintaining high perceived sarcasm expressivity (3.8 ± 0.2 SMOS). However, because its SMOS is comparable to the LoRA-only condition, we interpret the subjective results as evidence for complementary cue contributions rather than a clear additive perceptual advantage. This result aligns with prior perceptual studies showing that sarcasm recognition is facilitated when semantic content and intonation are congruent (Bryant, 2010; Bryant & Fox Tree, 2002; Woodland & Voyer, 2011). For instance, Woodland and Voyer (2011) found that incongruence between context and tone led to ambiguous or delayed judgments of sarcasm. Similarly, Bryant (2010) and Bryant and Fox Tree (2002) demonstrated that prosodic cues systematically signal sarcastic intent in spontaneous speech, particularly when linguistic cues alone are ambiguous.

Taken together, these results highlight three key insights: (i) raw, general-purpose embeddings are insufficient and can reduce perceived expressivity; (ii) parameter-efficient adaptation with LoRA can inject semantic and pragmatic knowledge that enhances the recognition of sarcasm; and (iii) retrieval-augmented conditioning further refines prosodic expressiveness by grounding synthesis in real sarcastic exemplars. In summary, these findings suggest the complementary roles of semantic and prosodic cues in sarcasm perception.

Conclusion

This work presents a unified computational framework for modeling sarcastic speech that captures how semantic and prosodic cues interact to convey sarcastic intent. Semantic representations are derived from a LoRA-adapted LLaMA 3, encoding discourse-level cues indicative of sarcasm, while prosodic exemplars retrieved via the RAG module provide patterns of intonation and emphasis characteristic of sarcastic delivery. By integrating these components, the framework enables the generation of speech that reflects the subtle interplay between meaning and prosody, producing utterances that are both intelligible and perceptually recognizable as sarcastic.

Beyond improvements in synthesis quality, our findings offer insights into the cognitive mechanisms underlying sarcasm perception. Results from both objective and subjective evaluations suggest that semantic and prosodic cues make complementary contributions to the recognition of sarcasm. This supports theoretical accounts of sarcasm as a pragmatic phenomenon arising from the interaction of meaning and delivery, rather than from isolated linguistic or acoustic features.

The proposed framework demonstrates how advances in LLMs and retrieval-based conditioning can be used to model subtle pragmatic phenomena. Unlike conventional emotional TTS systems that rely on coarse affective labels, our approach offers a more flexible and cognitively grounded way of controlling expressive speech through high-level semantic cues. Future work could further emphasize contextual information in sarcasm modeling, moving beyond isolated utterances toward discourse-aware detection and generation.

References

Akuzawa, K., Iwasawa, Y., & Matsuo, Y. (2018). Expressive speech synthesis via modeling expressions with variational autoencoder. *Proc. Interspeech 2018*, 3067–3071.

Bakhturina, E., Lavrukhin, V., Ginsburg, B., & Zhang, Y. (2021). Hi-fi multi-speaker english tts dataset. *Proc. Interspeech 2021*, 2776–2780.

Bryant, G. A. (2010). Prosodic contrasts in ironic speech. *Discourse Processes*, *47*(7), 545–566.

Bryant, G. A., & Fox Tree, J. E. (2002). Recognizing verbal irony in spontaneous speech. *Metaphor and symbol*, *17*(2), 99–119.

Cai, Y., Wu, Z., Jia, J., & Meng, H. (2024). Lora-mer: Low-rank adaptation of pre-trained speech models for multimodal emotion recognition using mutual information. *Proc. Interspeech 2024*, 4658–4662.

Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., & Poria, S. (2019, July). Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Acl* (pp. 4619–4629). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1455

Cheang, H. S., & Pell, M. D. (2008). The sound of sarcasm. *Speech communication*, *50*(5), 366–381.

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

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Feng, X., & Yoshimoto, A. (2024). Llama-vits: Enhancing tts synthesis with semantic awareness. *LREC-COLING 2024*, 10642–10656.

Gao, X., Bansal, S., Gowda, K., Li, Z., Nayak, S., Kumar, N., & Coler, M. (2024). Amused: An attentive deep neural network for multimodal sarcasm detection incorporating bi-modal data augmentation. *arXiv preprint arXiv:2412.10103*.

Kim, J., Kong, J., & Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *International Conference on Machine Learning*, 5530–5540.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*, *33*, 9459–9474.

Li, S., Mao, Q., & Shi, J. (2024). PL-TTS: A Generalizable Prompt-based Diffusion TTS Augmented by Large Language Model. *Interspeech 2024*, 4888–4892. https://doi.org/10.21437/Interspeech.2024-1429

Li, T., Yang, S., Xue, L., & Xie, L. (2021). Controllable emotion transfer for end-to-end speech synthesis. *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 1–5.

Li, Z., Gao, X., Nayak, S., & Coler, M. (2023). Sarcastic-speech: Speech synthesis for sarcasm in low-resource scenarios. *Proc. SSW 2023*, 242–243.

Li, Z., Gao, X., Zhang, Y., Nayak, S., & Coler, M. (2024). A functional trade-off between prosodic and semantic cues in conveying sarcasm. *Proc. Interspeech 2024*, 1070–1074.

Li, Z., Gao, X., Zhang, Y., Nayak, S., & Coler, M. (2025). Evaluating multimodal large language models on spoken sarcasm understanding. *arXiv preprint arXiv:2509.15476*.

Li, Z., Zhang, Y., Gao, X., Raghuvanshi, D., Kumar, N., Nayak, S., & Coler, M. (2025). Integrating feedback loss from bi-modal sarcasm detector for sarcastic speech synthesis. *Proc. SSW 2025*, 150–156.

Luo, D., Ma, C., Li, W., Wang, J., Chen, W., & Wu, Z. (2025, April). AutoStyle-TTS: Retrieval-Augmented Generation based Automatic Style Matching Text-to-Speech Synthesis [arXiv:2504.10309 [cs]]. https://doi.org/10.48550/arXiv.2504.10309

Misra, R., & Arora, P. (2023). Sarcasm detection using news headlines dataset. *AI Open*, *4*, 13–18. https://doi.org/https://doi.org/10.1016/j.aiopen.2023.01.001

Peters, S., & Almor, A. (2017). Creating the sound of sarcasm. *Journal of Language and Social Psychology*, *36*(2), 241–250.

Raghuvanshi, D., Gao, X., Li, Z., Bansal, S., Coler, M., Kumar, N., & Nayak, S. (2025). Intra-modal relation and emotional incongruity learning using graph attention networks for multimodal sarcasm detection. *ICASSP*, 1–5.

Ray, A., Mishra, S., Nunna, A., & Bhattacharyya, P. (2022). A multimodal corpus for emotion recognition in sarcasm. *LREC*, 6992–7003. https://aclanthology.org/2022.lrec-1.756

Ritschel, H., Aslan, I., Sedlbauer, D., & André, E. (2019). Irony man: Augmenting a social robot with the ability to use irony in multimodal communication with humans. *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 86–94.

Rockwell, P. (2000). Lower, slower, louder: Vocal cues of sarcasm. *Journal of Psycholinguistic research*, *29*(5), 483–495.

Rockwell, P. (2007). Vocal features of conversational sarcasm: A comparison of methods. *Journal of psycholinguistic research*, *36*(5), 361–369.

Shen, M., Zhang, S., Wu, J., Xiu, Z., AlBadawy, E., Lu, Y., Seltzer, M., & He, Q. (2025). Get large language models ready to speak: A late-fusion approach for speech generation. *ICASSP*, 1–5.

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. *Findings of EMNLP 2021*, 3784–3803.

Wang, Y., Stanton, D., Zhang, Y., Ryan, R.-S., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F., & Saurous, R. A. (2018). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *International conference on machine learning*, 5180–5189.

Woodland, J., & Voyer, D. (2011). Context and intonation in the perception of sarcasm. *Metaphor and Symbol*, *26*(3), 227–239.

Xue, J., Deng, Y., Gao, Y., & Li, Y. (2024). Retrieval Augmented Generation in Prompt-based Text-to-Speech Synthesis with Context-Aware Contrastive Language-Audio Pretraining. *Interspeech 2024*, 1800–1804. https://doi.org/10.21437/Interspeech.2024-980

Yadavalli, A., Pimentel, T., Regev, T. I., Wilcox, E., & Warstadt, A. (2025). What do prosody and text convey? characterizing how meaningful information is distributed across multiple channels. *arXiv preprint arXiv:2512.16832*.