ParaSpeechRAG: A Benchmark for Robust Speech Retrieval under Paralinguistic and Acoustic Perturbations

Anonymous ACL submission

Abstract

Speech Retrieval-Augmented Generation (Speech RAG) systems are increasingly used in real-world settings where spoken queries vary across speakers, emotions, and acoustic conditions. However, current evaluations primarily focus on clean-condition retrieval accuracy, leaving unclear whether these systems remain semantically invariant under paralinguistic variation. We introduce ParaSpeechRAG, a counterfactual robustness benchmark that isolates non-semantic acoustic factors while preserving lexical content. The benchmark spans controlled perturbations across speaker identity, emotion and prosody, additive noise, speaking rate, and codec compression, enabling systematic evaluation of semantic invariance in speech retrieval. We evaluate three retriever families: ASR-based cascades, speech-text multimodal retrievers (CLASP), and general-purpose audio-language models (GLAP). Results reveal a consistent robustness–accuracy trade-off: multimodal retrievers achieve higher performance under clean conditions but are substantially more sensitive to paralinguistic and acoustic perturbations, particularly emotion and noise. In contrast, ASR-based cascades are more robust to these variations due to their text-centric bottleneck, though with reduced expressive capacity. These findings highlight an overlooked dimension of instability in Speech RAG systems and underscore the need for robustness-aware evaluation beyond clean-condition benchmarks.

1 Introduction

Retrieval-Augmented Generation has fundamentally changed how language systems access knowledge, but when the query arrives as speech, a silent assumption underlies every retrieval step: that the system responds to *what* is said, not to *who says it* or *how*.

Unlike text queries, spoken utterances contain substantial paralinguistic (Lavan et al., 2019). The same sentence can produce markedly different acoustic realizations depending on speaker identity, emotional state, speaking style, recording conditions, or transmission artifacts, while preserving identical semantic content. As modern speech and audio embedding models necessarily encode these acoustic factors (Pasad et al., 2021; wen Yang et al., 2021), retrieval representations may vary even when meaning remains unchanged. If retrieval outcomes shift under such variations, then the system is not retrieving purely based on semantics.

Speech RAG systems typically rely either on ASR-based retrieval pipelines or on multimodal speech-text retrievers that directly map spoken queries into shared embedding spaces (Sidiropoulos and Kanoulas, 2024; Sun et al., 2025; Chen et al., 2025). While these approaches achieve strong retrieval performance under clean conditions, it remains unclear whether their representations are robust to non-semantic acoustic variation.

Existing benchmarks, including Spoken-SQuAD (Li et al., 2018), HeySQuAD (Wu et al., 2024), and SQuAD-SRC (Tang et al., 2023), evaluate *end-to-end QA accuracy* in relatively clean conditions. Evaluation of retrieval specifically, and of robustness to acoustic perturbations, remains an open gap. In the text domain, robustness benchmarks such as RARE (Zeng et al., 2025) and RGB (Chen et al., 2024) have revealed that even high-performing RAG systems are sensitive to query and document perturbations. No analogous study exists for the speech domain, where perturbations are qualitatively different: they are acoustic rather than lexical.

This gap is not merely academic. Real-world deployments of Speech RAG must cope with diverse speakers, emotional speech, background noise, lossy codecs, and spontaneous speaking styles.

A retriever that is sensitive to these factors will silently degrade, returning wrong passages without any signal to the downstream generator that retrieval has failed.

In this work, we address the following three research questions:

RQ1 *To what extent do current Speech RAG retrievers maintain semantic invariance under controlled paralinguistic perturbations?*

RQ2 *Which perturbation axes, and at what severity levels, most damage retrieval performance, and does the answer differ across retriever architectures (ASR cascade vs. speech-text multimodal vs. general audio-text)?*

RQ3 *Does the performance gap between multimodal retrievers and the ASR cascade observed in clean conditions reverse or narrow under paralinguistic perturbations, revealing a previously unrecognized robustness–accuracy trade-off?*

To answer these questions, we construct a systematic perturbation benchmark using CLASP (Abootorabi and Asgari, 2025) on top of Spoken-SQuAD and VoxPopuli, applying five categories of acoustic perturbation while holding semantic content constant. We evaluate three representative retriever families and report Recall@k and MRR across all conditions.

Our main contributions are:

1. The first benchmark for *paralinguistic sensitivity* in speech retrieval (ParaSpeechRAG), covering five perturbation axes with controlled severity levels.

2. A comparative evaluation of retriever architectures, revealing a fundamental robustness, accuracy trade-off that is invisible in clean-condition benchmarks.

3. Empirical evidence that multimodal audio embeddings encode paralinguistic information in ways that meaningfully harm retrieval, with practical implications for system design.

2 Related Works

The extension of RAG to spoken input has followed two main paradigms. The cascade paradigm, ASR followed by text retrieval, was the de facto standard in early spoken QA work (Li et al., 2018) and remains a competitive baseline due to the maturity of ASR systems. Spoken-SQuAD (Li et al., 2018) and HeySQuAD (Wu et al., 2024) established benchmarks for this setting, showing that ASR errors propagate into retrieval and that human-spoken queries carry variability that synthetic benchmarks underestimate. SQuAD-SRC (Tang et al., 2023) further extended this to multi-accent scenarios. ODSQA (Lee et al., 2018) provided an early real-speech alternative to TTS-generated datasets.

More recent work has sought to eliminate the ASR bottleneck by learning joint speech-text representations. Sidiropoulos and Kanoulas (2024) propose a multimodal dense retriever for open-domain QA from spoken queries, trained end-to-end to map speech directly into the retrieval space. SEAL (Sun et al., 2025) extends this to speech large language models, showing that embedding-based speech retrieval significantly reduces latency and error propagation compared to ASR cascades. WavRAG (Chen et al., 2025) introduces a dual-modality encoder that retrieves from both audio and text knowledge bases simultaneously, enabling truly multimodal document corpora.

The retrieval capabilities of multimodal audio systems have advanced substantially through contrastive pretraining. CLAP (Elizalde et al., 2023) learns a joint audio-text embedding space through contrastive objectives over large-scale audio-caption pairs, achieving strong zero-shot performance on audio classification and text-to-audio retrieval. Wu et al. (2023) extend CLAP with feature fusion and keyword-to-caption augmentation, further improving retrieval on the LAION-Audio-630K benchmark. T-CLAP (Yuan et al., 2024) addresses temporal reasoning limitations in CLAP by introducing temporal-contrastive captions and a specialized loss. CoLLAP (Wu et al., 2025) scales contrastive audio-language pretraining to long-form audio, improving retrieval over extended acoustic contexts.

While these models were originally designed for audio event retrieval and classification, their zero-shot cross-modal capabilities make them attractive as Speech RAG retrievers, a use case their training objectives did not anticipate.

The fragility of RAG systems under perturbation has received increasing attention in the text domain. Zeng et al. (2025) introduce RARE, a unified benchmark that stress-tests query and document perturbations over dynamic corpora, finding

that RAG systems are systematically sensitive to input noise. Perçin et al. (2025) study robustness to typographical and linguistic variations at the retrieval stage, showing that mid-sized generators can outperform larger ones under perturbation. RGB (Chen et al., 2024) evaluates four capacity dimensions, noise robustness, negative rejection, information integration, and counterfactual resistance, revealing consistent weaknesses in LLMs handling distracting context.

None of these works considers acoustic perturbation, and none evaluates systems where the query modality is speech. Our work fills this gap by designing an analogous robustness protocol specifically for the acoustic domain.

The robustness of speech representations under acoustic perturbation has been extensively studied in ASR and speaker verification, though rarely in retrieval settings. Prior work shows that noise, reverberation, and channel degradation can substantially alter downstream speech representations (Tseng and Harwath, 2025). Meanwhile, modern voice conversion systems (Choi et al., 2023; Li et al., 2023) and expressive speech synthesis frameworks enable controlled manipulation of speaker identity and emotional prosody while preserving lexical content, making them suitable tools for counterfactual robustness evaluation.

3 Methodology

3.1 Task Formulation

We study the *retrieval* stage of Speech RAG in isolation. Given a spoken query q (an audio signal) and a text corpus $\mathcal{C} = \{p_1, \ldots, p_N\}$ of passages, a retriever f returns a ranked list of passages. Retrieval is correct if the ground-truth evidence passage p^* appears in the top-k results.

We evaluate robustness by applying a paralinguistic perturbation ϕ to the query audio, producing $\tilde{q} = \phi(q)$, and measuring how much retrieval performance degrades relative to the clean condition. By construction, $\tilde{q}$ and q have *identical semantic content*, only the acoustic surface form changes. A retriever that is paralinguistically invariant should maintain its clean-condition performance under all ϕ.

3.2 Retriever Systems

We evaluate three retriever families that together span the design space of Speech RAG:

ASR Cascade (Parakeet TDT + Text Retrieval). The spoken query is transcribed by Parakeet TDT (Xu et al., 2023), a state-of-the-art ASR model. The transcript is then used to retrieve passages using LaBSE (Feng et al., 2022). This system is our *traditional baseline*: it is the most widely deployed approach and benefits from the maturity and robustness of modern text retrieval pipelines. Its main limitation is error propagation from the ASR stage, particularly under adverse acoustic conditions.

CLASP (Speech-Text Multimodal Retriever). CLASP is trained to align speech and text representations in a shared embedding space, enabling direct retrieval of text passages from spoken queries without ASR. We evaluate two pooling variants: *mean pooling* (CLASP-mean), which aggregates over the full audio sequence, and *chunked max-sim* (CLASP-chunked), which splits longer utterances into chunks and uses maximum similarity across chunks, inspired by ColBERT-style late interaction.

GLAP (General Audio-Text Retriever). GLAP is a general-purpose audio-language model trained on diverse audio events, music, and speech. Unlike CLASP, it was not specifically trained for spoken QA retrieval. We include it as a probe of zero-shot cross-modal retrieval from a model with broader acoustic coverage but less speech-specific alignment.

3.3 Perturbation Axes

All perturbations are applied to the query audio only, keeping the corpus unchanged. For each axis, we define multiple severity levels to trace performance degradation curves.

Speaker Variation (ϕ_{spk}). We apply zero-shot voice conversion to replace the original speaker identity in Spoken-SQuAD dev queries while preserving phonetic content. Each query is converted to six target voices: four speakers drawn from LibriSpeech (Panayotov et al., 2015) and two reference voices (Azuma and Trump), totalizing three male and three female targets. To ensure robustness of our findings across conversion systems, we generate two parallel sets of converted audio using Seed-VC (Liu, 2024) and GenVC (Cai et al., 2025), two independent zero-shot voice conversion models. This set spans diverse vocal characteristics, allowing us to test whether retriever embeddings are invariant to speaker identity.

Emotion and Prosody (ϕ_{emo}). We re-synthesize the Spoken-SQuAD dev queries via zero-shot voice conversion using Seed-VC (Liu, 2024), conditioned on expressive reference utterances from Actor 01 of the RAVDESS dataset (Livingstone and Russo, 2018). We cover four discrete emotion categories (neutral, happy, sad, angry) at two intensity levels (normal, strong), yielding seven condition combinations. The intensity level serves as a proxy for acoustic activation, allowing us to disentangle the effect of emotional category from the effect of expressive magnitude on retrieval performance.

Additive Noise (ϕ_{noise}). Following Tseng and Harwath (2025), we add white noise and convolutive reverberation at four SNR levels: 20, 15, 10, and $\sim$5 dB. These cover the range from mild environmental noise to severely degraded speech.

Speaking Rate (ϕ_{rate}). We apply time-scale modification to produce speech at $0.5\times$, $0.625\times$, $0.75\times$, $0.875\times$, $1.25\times$, $1.5\times$, $1.75\times$ and $2.0\times$ the original rate without affecting pitch. This perturbation is orthogonal to content and tests sensitivity to temporal structure in audio embeddings.

Codec Compression (ϕ_{codec}). We encode and decode audio through MP3 at bitrates ranging from 8 to 128 kbps, and through Opus at equivalent quality levels. This simulates transmission over bandwidth-constrained channels.

3.4 Evaluation Metrics

We report Recall@k (R@k) for $k \in \{1, 5, 10\}$ and Mean Reciprocal Rank (MRR); mean and median rank are additionally recorded for diagnostic purposes.[1] For each perturbation condition we additionally report the relative degradation ΔR@1 with respect to the clean baseline. All results are computed over the full Spoken-SQuAD validation set (2,067 paragraphs); each query is ranked against the entire validation pool.

4 Experimental Setup

4.1 Training Details

We train CLASP on the concatenation of Spoken-SQuAD and VoxPopuli English, which provide complementary supervision signals for Speech RAG retrieval. Spoken-SQuAD provides task-aligned spoken question answering pairs that closely match the downstream retrieval setting evaluated in our benchmark. Additionally, its comparatively controlled acoustic conditions make it particularly suitable for generating counterfactual perturbations, since performance variations can be more reliably attributed to the applied acoustic transformations rather than uncontrolled environmental factors. In contrast, VoxPopuli contributes substantially greater variability in speakers, recording conditions, speaking styles, and prosody through large-scale real-world parliamentary speech. Although VoxPopuli is not directly used in the perturbation benchmark itself, its acoustic diversity exposes the retriever to broader real-world variability during training, encouraging more robust and generalizable speech-text representations.

The final training set contains 18,896 spoken-query/text pairs from Spoken-SQuAD and 184,219 speech-text pairs from VoxPopuli English.

CLASP is trained using a contrastive objective over paired speech and text representations, enabling direct alignment between spoken queries and textual passages in a shared embedding space.

Training is performed for 50 epochs using the Adam optimizer with learning rate 1×10^{-4}, batch size 32. Model selection is based on validation contrastive loss computed on the Spoken-SQuAD development set.

4.2 Noise Robustness Evaluation

We evaluate retrieval robustness under three classes of acoustic corruption: white noise, ambient environmental noise, and reverberation. Following Tseng and Harwath (2025), perturbations are applied exclusively to the query audio while keeping the retrieval corpus unchanged, isolating the sensitivity of the query representation to non-semantic acoustic variation. White Gaussian noise and ESC-50 environmental recordings (Piczak) are added at 20, 10, and 5 dB SNR levels, corresponding to mild, moderate, and severe degradation. Reverberation is simulated through delayed room reflections at equivalent severity levels. All perturbed audio is processed through the same preprocessing pipeline as clean speech before retrieval evaluation.

4.3 Codec Compression

We evaluate retrieval robustness under lossy codec compression using **MP3** and **Opus** encoding. Perturbed queries are generated through an encode-decode pipeline implemented with `FFmpeg`, where each spoken query is compressed at a target bitrate

[1] Each query has a single relevant paragraph, so Recall@k is equivalent to the Hits@k implemented in our evaluation code, and Mean Average Precision coincides with MRR.

and reconstructed back to uncompressed PCM before retrieval. We evaluate five bitrate levels (8, 16, 32, 64, and 128 kbps) for both codec families, covering the range from severe compression to near-transparent quality. This perturbation tests whether retrieval embeddings remain stable under bandwidth-constrained transmission artefacts while preserving lexical content and utterance duration.

4.4 Speaker Variation

We evaluate retrieval robustness under speaker identity variation using zero-shot voice conversion. Spoken-SQuAD queries are converted to six target voices while preserving phonetic content, allowing us to isolate the sensitivity of retrieval embeddings to speaker-dependent acoustic characteristics. To reduce dependence on a single conversion framework, we generate parallel perturbation sets using two independent systems: Seed-VC (Liu, 2024) and GenVC (Cai et al., 2025). The target speakers include four LibriSpeech voices (Panayotov et al., 2015) (IDs 1089, 2803, 3081, and 6319) alongside two additional Seed-VC reference voices (*Azuma* and *Trump*), spanning both male and female speakers.

4.5 Emotion and Prosody

We evaluate retrieval robustness under emotion-driven prosodic variation using Seed-VC v2 (Liu, 2024). Spoken queries are re-synthesized using expressive reference utterances from the RAVDESS corpus (Livingstone and Russo, 2018), modifying pitch contour, vocal quality, rhythm, and spectral energy while preserving lexical content. We evaluate seven emotion conditions spanning four categories and two intensity levels: *neutral*, *happy* (normal, strong), *sad* (normal, strong), and *angry* (normal, strong). A single reference utterance from Actor 01 is used as the affective template for each condition. All converted audio is resampled to 16 kHz and processed through the same retrieval pipeline as clean speech.

4.6 Speaking Rate

We evaluate retrieval robustness under temporal speech variation using pitch-preserving time-scale modification (TSM). Spoken queries are stretched or compressed in time while preserving pitch and speaker identity, isolating the sensitivity of retrieval embeddings to speaking-rate variation independent of lexical content. Following Speech Robust Bench (Shah et al., 2024), we evaluate eight speed factors spanning both slow and fast speech: 0.5×, 0.625×, 0.75×, 0.875×, 1.25×, 1.5×, 1.75×, and 2.0×. This perturbation probes whether retrieval representations remain stable under substantial changes in temporal speech dynamics.

5 Results and Discussion

5.1 Clean Retrieval Performance

Table 1 reports retrieval performance for all systems on the clean Spoken-SQuAD validation set (2,067 paragraphs). CLASP with mean pooling achieves the highest results across all metrics, reaching a Recall@1 of 0.954 and an MRR of 0.971, with perfect Recall@5 and Recall@10. These results establish CLASP-mean as the clear upper bound within the evaluated retriever families and confirm that direct speech-text alignment, when pooled globally, produces highly discriminative query representations.

The ASR + LaBSE cascade achieves a Recall@1 of 0.694 and an MRR of 0.771, positioning itself as a competitive baseline given its reliance on transcription as an intermediate step. This performance level demonstrates the maturity of ASR-then-retrieve pipelines under clean acoustic conditions, while also establishing the ceiling that transcription errors impose.

In contrast, the CLASP-chunked (max-sim) variant reaches only 0.547 Recall@1 (MRR 0.671), substantially below both CLASP-mean and the ASR cascade. This gap suggests that chunked ColBERT-style late interaction pooling, while beneficial in text-based retrieval, may not straightforwardly transfer to speech. One possible explanation is that segmenting speech into fixed-size chunks introduces boundary artifacts and mismatches between the temporal granularity of audio and text representations, penalizing retrieval for short queries that derive limited benefit from chunk-level aggregation.

GLAP, despite being a general audio-language model, achieves moderate zero-shot performance (Recall1 0.209, MRR 0.289). Although GLAP was not trained for spoken QA, its contrastive pretraining on diverse audio-caption pairs transfers non-trivially to this domain, suggesting a degree of semantic alignment between speech and text even without task-specific fine-tuning. CLAP, in turn, essentially fails on this task (Recall@1 $\approx$ 0.000), indicating that audio-event-oriented pretraining alone is insufficient for speech-based fac-

Table 1: Clean retrieval on Spoken-SQuAD validation (2,067 paragraphs).

System	Recall@1	Recall@5	Recall@10	MRR
CLASP (mean)	**0.954**	**1.0**	**1.0**	**0.971**
CLASP (chunked, max-sim)	0.547	0.827	0.898	0.671
CLAP (audio-only)	0.000	0.001	0.002	0.001
GLAP (audio-only)	0.209	0.366	0.442	0.289
ASR + LaBSE (cascade)	0.694	0.868	0.911	0.771

tual retrieval, where the retrieval signal is lexical rather than acoustic-event-based.

5.2 Paralinguistic Robustness

Table 2: Results varying the target voice in the conversion (Mean of Models), comparing the proposed approach against the ASR Cascade baseline. All values are presented as **Recall@1 / MRR**.

Target Voice	Gender	Cascade	CLASP
1089	Male	0.695 / 0.772	**0.919 / 0.950**
2803	Male	0.696 / 0.772	0.908 / 0.943
trump	Male	0.688 / 0.766	0.890 / 0.930
3081	Female	0.694 / 0.771	0.818 / 0.881
6319	Female	0.694 / 0.771	0.916 / 0.948
azuma	Female	0.693 / 0.770	0.884 / 0.928
WAV (Original/Baseline)		*0.694 / 0.771*	*0.954 / 0.972*
Mean (Converted)		*0.693 / 0.770*	*0.889 / 0.930*

Table 2 presents retrieval performance under speaker variation, evaluated using zero-shot voice conversion across six target identities (three male, three female). Two key findings emerge from these results.

First, the ASR cascade exhibits near-perfect invariance to speaker identity: its Recall@1 and MRR remain essentially constant across all converted voices (mean 0.693/0.770), indistinguishable from the original audio baseline (0.694/0.771). This stability is expected, since speaker-related acoustic cues are largely removed during ASR transcription before retrieval. Speaker variation affects the cascade only through potential ASR degradation, which, under these controlled conditions with a high-quality voice conversion model, appears negligible.

Second, CLASP-mean, while substantially outperforming the cascade in clean conditions, is measurably sensitive to speaker identity: mean Recall@1 drops from 0.954 (baseline) to 0.889 ($\Delta = -6.8\%$), and MRR from 0.972 to 0.930 ($\Delta = -4.3\%$). This degradation is not uniform across target voices. Cross-gender conversion to voice 3081 (female) induces the largest drop (Recall@1 = 0.818), while same-gender and other cross-gender targets cause smaller but still consistent drops. The pattern suggests that CLASP's speech encoder encodes speaker-correlated acoustic features, that shift query embeddings away from the representations computed during training, even when lexical content is preserved.

These results reveal a robustness–accuracy tradeoff that is largely hidden under clean-condition evaluation: although CLASP-mean achieves substantially higher retrieval performance, the ASR cascade remains markedly more stable under speaker variation due to the invariance introduced by the acoustic-to-text bottleneck. Consequently, multimodal retrievers may degrade as speaker characteristics shift, even when semantic content remains unchanged.

Table 3: Results across different emotional states and intensities, comparing the Multimodal approach (CLASP) against the ASR Cascade baseline. All values are presented as **Recall@1 / MRR**.

Emotion	Intensity	Cascade	CLASP
Angry	Normal	0.695 / 0.771	0.769 / 0.848
	Strong	0.679 / 0.756	0.672 / 0.771
Happy	Normal	0.694 / 0.771	**0.836 / 0.894**
	Strong	0.693 / 0.769	0.651 / 0.757
Neutral	Normal	0.695 / 0.771	0.799 / 0.868
Sad	Normal	0.697 / 0.773	0.780 / 0.856
	Strong	0.692 / 0.769	0.770 / 0.847

Table 3 presents retrieval performance under emotion-driven prosodic variation. Compared to the speaker conversion setting, these perturbations introduce substantially stronger expressive variation while preserving semantic content, providing a more challenging test of robustness to paralinguis-

tic variation.

The ASR cascade remains comparatively stable across all emotional conditions, with Recall@1 varying only marginally from 0.697 (sad-normal) to 0.679 (strong anger). This behavior is consistent with the invariance introduced by the ASR bottleneck, which largely removes paralinguistic information before retrieval.

In contrast, CLASP exhibits substantial degradation under emotional perturbations despite achieving the strongest clean-condition performance. Relative to the clean baseline (Recall@1 = 0.954, MRR = 0.972), all emotional conditions reduce retrieval quality, with the strongest degradation occurring under high-intensity speech. Strong happy speech reduces Recall@1 to 0.651 (MRR = 0.757), while strong anger reaches 0.672 (MRR = 0.771), bringing CLASP close to the performance level of the ASR cascade.

Compared to speaker variation alone, emotional perturbations produce substantially larger degradation, suggesting that CLASP is particularly sensitive to expressive prosodic variation. Moreover, degradation consistently increases with emotional intensity: for both happy and angry speech, the strong condition performs markedly worse than the normal condition despite identical lexical content. This pattern suggests that retrieval instability is driven primarily by the magnitude of acoustic deviation from the training distribution rather than by emotion category itself.

These results indicate that emotional and prosodic characteristics remain partially entangled with semantic representations in multimodal speech retrieval embeddings, leading to retrieval degradation even when semantic content is preserved.

5.3 Acoustic Perturbation Robustness

Figure 1 summarizes the robustness behavior of CLASP under four classes of acoustic perturbation: codec compression, speaking-rate modification, additive noise, and reverberation.

A first observation is that the degradation profiles differ substantially across perturbation axes, indicating that not all non-semantic acoustic variations affect retrieval equally. While some perturbations produce only minor shifts in retrieval quality, others cause catastrophic collapse despite preserving the underlying lexical content. This asymmetry reinforces the central premise of our benchmark: Speech RAG retrievers are not uniformly invariant to paralinguistic and acoustic transformations.

The codec compression results reveal that CLASP is comparatively robust to bandwidth-related degradation. Both MP3 and Opus maintain retrieval performance close to the clean baseline across nearly the entire bitrate spectrum. Even under severe MP3 compression at 8 kbps, Recall@1 remains at 0.885 and MRR at 0.927, representing only moderate degradation relative to the clean condition (0.954 / 0.972). Opus exhibits even stronger robustness, remaining near-baseline across all evaluated bitrates, including 8 kbps. This behavior suggests that codec artifacts primarily remove perceptually redundant spectral detail while preserving the coarse phonetic and temporal structure required by the retriever.

In contrast, the speaking-rate experiments demonstrate that CLASP is substantially more sensitive to temporal distortions than to compression artifacts. Retrieval quality degrades progressively as playback speed diverges from the original utterance. Moderate changes in speaking rate remain relatively tolerable, but extreme accelerations and slowdowns cause pronounced collapse. At 0.5× speed, Recall@1 falls to 0.674, while at 2.0× it further drops to 0.630. The asymmetry between mild and extreme modifications suggests that CLASP partially relies on temporal dynamics learned during training, including rhythm, phoneme duration, and prosodic pacing. When speech deviates substantially from canonical temporal structure, the resulting embeddings drift away from their original semantic neighborhood.

The most severe degradation appears under additive noise and environmental corruption. All three noise families produce dramatic reductions relative to the clean baseline, with ambient noise causing the strongest collapse overall. Under ambient noise at 5 dB, Recall@1 drops to 0.301 and MRR to 0.389, corresponding to more than a 65% relative reduction in top-1 retrieval accuracy. White noise produces similarly severe degradation, falling from 0.954 Recall@1 in the clean setting to 0.335 at 5 dB. These results indicate that the retrieval representation is highly vulnerable to acoustic corruption that alters spectral consistency and phonetic salience.

Interestingly, reverberation exhibits a qualitatively different behavior from additive noise. While performance remains substantially below the clean baseline across all reverberation conditions, degradation does not monotonically worsen with de-

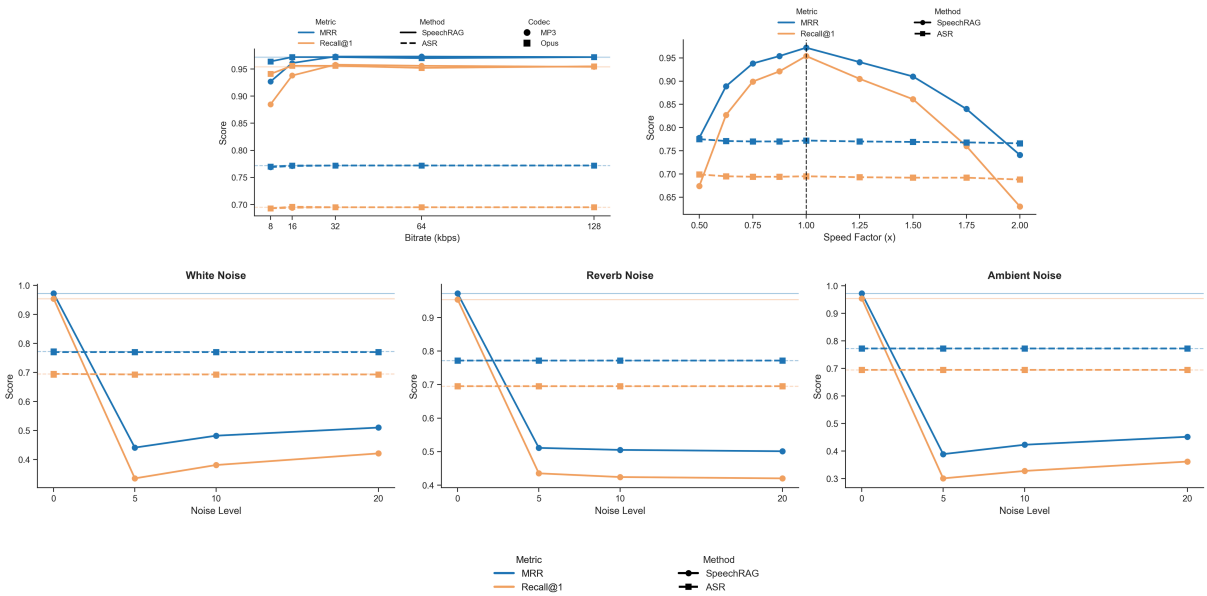

Figure 1: It describes the 5 subframes (codec, rate, white noise, reverb, ambient), identifies solid lines (MRR) vs. dashed lines (Recall@1), maps CLASP vs. ASR, and explains the x-axis orientation in each subframe.

creasing SNR. Recall@1 remains relatively stable between 0.420 and 0.435 across all reverberation severities. Unlike white or ambient noise, which inject competing acoustic energy into the signal, reverberation preserves much of the underlying speech content while redistributing it temporally. Consequently, retrieval degradation saturates early instead of collapsing progressively with severity.

These analyses reveal a clear robustness hierarchy: codec compression causes limited degradation, while speaking-rate variation, additive noise, and reverberation produce substantially larger retrieval instability. The results suggest that CLASP is comparatively robust to perceptual compression artifacts but remains sensitive to perturbations that distort temporal and prosodic speech structure, likely reflecting the predominance of clean and acoustically canonical speech during training.

6 Conclusion

In this work, we introduced ParaSpeechRAG, the first benchmark for evaluating paralinguistic robustness and semantic invariance in Speech Retrieval-Augmented Generation systems. Unlike prior work focused on clean-condition retrieval or ASR robustness, our benchmark isolates non-semantic acoustic variations while preserving lexical content, enabling controlled analysis of sensitivity to speaker identity, emotion, speaking rate, codec compression, reverberation, and noise.

Our experiments reveal a consistent robustness–accuracy trade-off across Speech RAG architectures. While multimodal speech-text retrievers such as CLASP achieve strong performance under clean conditions, they are substantially more sensitive to paralinguistic and acoustic perturbations than ASR-based cascades. In particular, emotional prosody, speaking-rate variation, and additive noise induce significant retrieval instability despite preserving semantics, indicating that current speech retrieval embeddings remain partially entangled with non-semantic acoustic factors.

These results show that clean-condition evaluation is insufficient for real-world Speech RAG systems, and that robustness to acoustic and paralinguistic variation should be treated as a core evaluation dimension alongside retrieval accuracy and latency.

Limitations

Our study has several limitations that should be considered when interpreting the results. Although our perturbations are carefully controlled, including TTS-generated audio, synthetic noise addition, and voice conversion, they may not fully reflect naturally occurring acoustic degradation in real-world environments, which limits the ecological validity of our robustness analysis. In addition, our benchmark is currently restricted to English datasets (Spoken-SQuAD and VoxPopuli English),

which may constrain generalization to multilingual settings, particularly for languages with richer prosodic variation and different phonetic structures. Furthermore, our evaluation focuses exclusively on the retrieval stage of Speech RAG pipelines, and therefore does not capture how retrieval instability propagates to downstream generation quality, leaving end-to-end behavior under paralinguistic variation as an open question. Finally, we do not explicitly model conversational dynamics such as speech spontaneity or variable utterance duration, both of which may influence robustness in more natural interactive settings.

References

Mohammad Mahdi Abootorabi and Ehsaneddin Asgari. 2025. Clasp: Contrastive language-speech pretraining for multilingual multimodal information retrieval. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part IV*, page 10–20, Berlin, Heidelberg. Springer-Verlag.

Zexin Cai, Henry Li Xinyuan, Ashi Garg, Leibny Paola García-Perera, Kevin Duh, Sanjeev Khudanpur, Matthew Wiesner, and Nicholas Andrews. 2025. Genvc: Self-supervised zero-shot voice conversion. *Preprint*, arXiv:2502.04519.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762.

Yifu Chen, Shengpeng Ji, Haoxiao Wang, Ziqing Wang, Siyu Chen, Jinzheng He, Jin Xu, and Zhou Zhao. 2025. WavRAG: Audio-integrated retrieval augmented generation for spoken dialogue models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12505–12523, Vienna, Austria. Association for Computational Linguistics.

Ha-Yeong Choi, Sang-Hoon Lee, and Seong-Whan Lee. 2023. Diff-HierVC: Diffusion-based Hierarchical Voice Conversion with Robust Pitch Generation and Masked Prior for Zero-shot Speaker Adaptation. In *Interspeech 2023*, pages 2283–2287.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

N. Lavan, A. M. Burton, S. K. Scott, and C. McGettigan. 2019. Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review*, 26(1):90–102.

Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung-Yi Lee. 2018. Odsqa: Open-domain spoken question answering dataset. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 949–956.

Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung yi Lee. 2018. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. *Preprint*, arXiv:1804.00320.

Jingyi Li, Weiping Tu, and Li Xiao. 2023. Freevc: Towards high-quality text-free one-shot voice conversion. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Songting Liu. 2024. Zero-shot voice conversion with diffusion transformers. *Preprint*, arXiv:2411.09943.

Steven R. Livingstone and Frank A. Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):1–35.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference (SciPy)*, pages 18–25.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

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

Sezen Perçin, Xin Su, Qutub Sha Syed, Phillip Howard, Aleksei Kuvshinov, Leo Schwinn, and Kay-Ulrich Scholl. 2025. Investigating the robustness of retrieval-augmented generation at the query level. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 439–457, Vienna, Austria. Association for Computational Linguistics.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.

Muhammad A. Shah, David Solans Noguero, Mikko A. Heikkila, Bhiksha Raj, and Nicolas Kourtellis. 2024. Speech robust bench: A robustness benchmark for speech recognition. *Preprint*, arXiv:2403.07937.

Georgios Sidiropoulos and Evangelos Kanoulas. 2024. A multimodal dense retrieval approach for speech-based open-domain question answering. *Preprint*, arXiv:2409.13483.

Chunyu Sun, Bingyu Liu, Zhichao Cui, Junhan Shi, Anbin Qi, Tian hao Zhang, Dinghao Zhou, and Lewei Lu. 2025. Seal: Speech embedding alignment learning for speech large language model with retrieval-augmented generation. *Preprint*, arXiv:2502.02603.

Yixuan Tang, Anthony KH Tung, and Edith Elkind. 2023. Squad-src: A dataset for multi-accent spoken reading comprehension. In *IJCAI*, pages 5206–5214.

Wei-Cheng Tseng and David Harwath. 2025. Probing the robustness properties of neural speech codecs. *Preprint*, arXiv:2505.24248.

Jean-Marc Valin, Koen Vos, and Timothy Terriberry. 2012. Definition of the Opus Audio Codec. RFC 6716, IETF.

Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Interspeech 2021*, pages 1194–1198.

Junda Wu, Warren Li, Zachary Novack, Amit Namburi, Carol Chen, and Julian McAuley. 2025. Collap: Contrastive long-form language-audio pretraining with musical temporal structure augmentation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yijing Wu, SaiKrishna Rallabandi, Ravisutha Srinivasamurthy, Parag Pravin Dakle, Alolika Gon, and Preethi Raghavan. 2024. Heysquad: A spoken question answering dataset. *Preprint*, arXiv:2304.13689.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Hainan Xu, Fei Jia, Somshubra Majumdar, He Huang, Shinji Watanabe, and Boris Ginsburg. 2023. Efficient sequence transduction by jointly predicting tokens and durations. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Yi Yuan, Zhuo Chen, Xubo Liu, Haohe Liu, Xuenan Xu, Dongya Jia, Yuanzhe Chen, Mark D. Plumbley, and Wenwu Wang. 2024. T-clap: Temporal-enhanced contrastive language-audio pretraining. In *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6.

Yixiao Zeng, Tianyu Cao, Danqing Wang, Xinran Zhao, Zimeng Qiu, Morteza Ziyadi, Tongshuang Wu, and Lei Li. 2025. Rare: Retrieval-aware robustness evaluation for retrieval-augmented generation systems. *Preprint*, arXiv:2506.00789.

A Voice Conversion and Emotion Transfer Details

Speaker perturbations are generated using two independent zero-shot voice conversion systems: Seed-VC (Liu, 2024) and GenVC (Cai et al., 2025). Both systems transfer speaker identity from a reference utterance while preserving the linguistic content of the source speech.

For Seed-VC, inference uses 30 diffusion steps, CFG rate of 0.7, and length adjustment factor of 1.0. All converted audio is resampled to 16 kHz prior to embedding extraction.

For GenVC, we use the `GenVC_large` checkpoint with top-k sampling ($k = 15$). Source and reference audio are loaded at the model-specific sampling rates and converted output is evaluated without additional resampling.

Emotion conversion is performed with Seed-VC v2 (Liu, 2024), which disentangles linguistic content from speaker style via a two-stage pipeline: an autoregressive model extracts style tokens from the reference - encoding timbre, pitch patterns, and energy contour - while a flow-matching diffusion model (CFM) resynthesises the audio conditioned on those tokens alongside a content representation derived from the source. Operating in full style-transfer mode (convert_style=True) causes both the speaker characteristics and the affective prosodic patterns of the reference to be imposed on the source utterance, as opposed to identity-only conversion used in the speaker variation experiments.

B Codec Compression Details

Codec perturbations are implemented through an encode-decode reconstruction pipeline using `FFmpeg`. MP3 compression uses `libmp3lame`, while Opus compression uses `libopus` (Valin et al., 2012). Each spoken query is compressed at target bitrates of 8, 16, 32, 64, and 128 kbps before re-

construction to uncompressed PCM for retrieval evaluation.

For MP3, decoded audio is restored to PCM WAV format at the original sampling rate. For Opus, reconstructed waveforms are resampled to 24 kHz prior to retrieval evaluation to match the original audio configuration.

C Speaking Rate Perturbation Details

Speaking-rate perturbations are generated using pitch-preserving time-scale modification implemented through `librosa.effects.time_stretch` (McFee et al., 2015). The phase vocoder algorithm stretches or compresses speech temporally while preserving pitch and spectral characteristics.

We evaluate four slow-speech factors (0.5×, 0.625×, 0.75×, 0.875×) and four fast-speech factors (1.25×, 1.5×, 1.75×, and 2.0×), following the perturbation levels defined in Speech Robust Bench (Shah et al., 2024). Audio is processed at its original sampling rate without additional resampling.