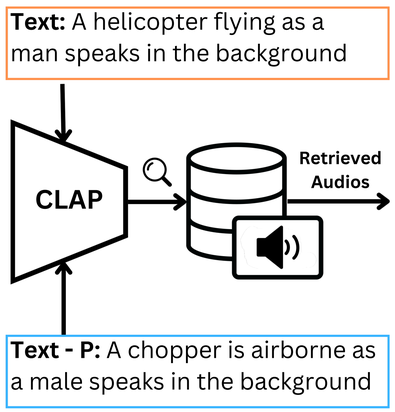
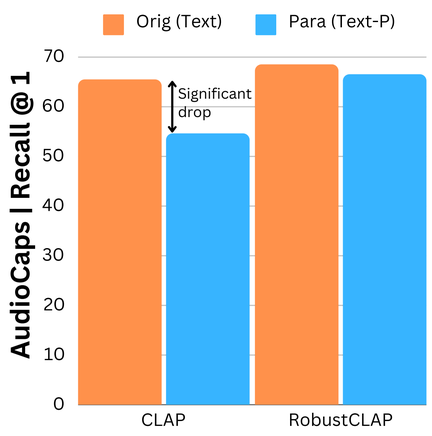
Do Audio-Language Models Understand Linguistic Variations?

Ramaneswaran Selvakumar*, Sonal Kumar*, Hemant Kumar Giri*
Nishit Anand, Ashish Seth, Sreyan Gosh, Dinesh Manocha
University of Maryland, College Park

Motivation

- Audio-Language Models (ALMs) like CLAP allow text-based retrieval of audio.
- Natural language queries are diverse and vary in syntax/wording.
- Human paraphrasing is natural — but can ALMs generalize to such variation?
- Early results: Up to 16% drop in retrieval when queries are paraphrased.





Problem with Existing ALMs

- Paraphrased queries result in significant drops in R@10.
- ALMs rely too much on lexical overlap, not semantics.
- Real-world users won't always use the exact training phrasing.

Introducing RobustCLAP

- Learns to align semantically equivalent paraphrases.
- Uses a multi-view contrastive loss with paraphrased training data.
- Requires minimal compute and data to train.

Data Synthesis Pipeline

Two-step LLM generation

- **Paraphrase Generation:** prompt LLM with human written ICL examples.
- **Paraphrase Correction:** prompt LLM to reason about correctness of paraphrase and correct if required

Sample caption, paraphrase and corrected paraphrase

Text: A person talking which later imitates a couple of meow sounds.
Text-P': An individual speaks, subsequently mimicking some cat cries.
Text-P: An individual speaks, subsequently mimicking some cat meows.

Multiview CLAP Loss

- CLAP loss reformulated to incorporate paraphrases
- T_{p1}: structural variation only
- T_{p2}: structure + vocab variation

$$L_{final} = L_{\text{clap}} + L^{\text{p}_1} + L^{\text{p}_2}$$

$$L^{\text{A}}_{\text{p}_k} = \sum_i \left[-\log \left(\frac{S(T_i^{\text{p}_k}, A_i)}{\sum_j S(T_i^{\text{p}_k}, A_j)} \right) \right]$$

$$L^{T}_{\text{p}_k} = \sum_i \left[-\log \left(\frac{S(T_i^{\text{p}_k}, T_i)}{\sum_j S(T_i^{\text{p}_k}, T_j)} \right) \right]$$

Main Retrieval Results

Benchmark →	AudioCaps		Clotho		Audioset SL		SoundDesc		DCASE	
Model ↓	TEST	TEST-P	TEST	TEST-P	TEST	TEST-P	TEST	TEST-P	TEST	TEST-P
ML-ACT	35.53	34.87	27.54	23.90	21.52	17.91	08.72	06.06	10.12	08.77
MSCLAP-22	84.74	84.63	86.74	43.94	27.73	23.72	14.33	11.87	39.91	30.99
MSCLAP-23	80.77	77.63	51.14	42.12	55.12	39.15	38.27	24.89	47.84	39.21
CompA	81.60	79.50	51.28	42.49	43.03	40.24	33.32	23.56	49.54	39.51
LAION-CLAP	82.86	82.84	52.03	43.98	46.91	41.94	24.62	18.09	44.73	37.81
RobustCLAP	85.62	81.55	57.27	53.47	57.44	53.64	25.48	21.54	54.66	50.35

Further Results

Impact Of Sound Event And Attributes

- Paraphrasing sound attributes -> 3.8% drop
- Paraphrasing sound events → 15% drop

Sound Classification Performance

- RobustCLAP does not have catastrophic forgetting of pretrained knowledge
- Sound classification performance is retained

Dataset	Model	
	CLAP	RobustCLAP
AudioCaps	65.51	68.54
+ Sound attributes mod.	61.96	68.12
+ Sound events mod.	50.24	65.48

Model	ESC-50	FSD-50K
CLAP	94.25	53.20
RobustCLAP	94.07	52.81

Error Analysis

Common Error Patterns Observed

- Spurious correlation to non-paraphrased sound events
- Captures the dominant context but lacks precision
- Does not capture sound attributes

Whats Next?

- Currently limited to short audio, need to expand to longer audios
- Paraphrasing quality depends on LLM, incorporate human supervision
- Captures the dominant context but lacks precision
- Limited to sound, expand to music and speech