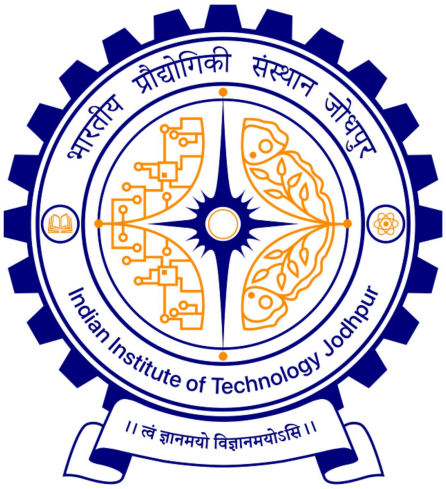
A REPORT

On

ASSIGNMENT 3(Speech Understanding)

Shripad pate(M23MAC007)



Data and Computational Sciences (DCS)
Department of Mathematics
Indian Institute of Technology, Jodhpur

2025

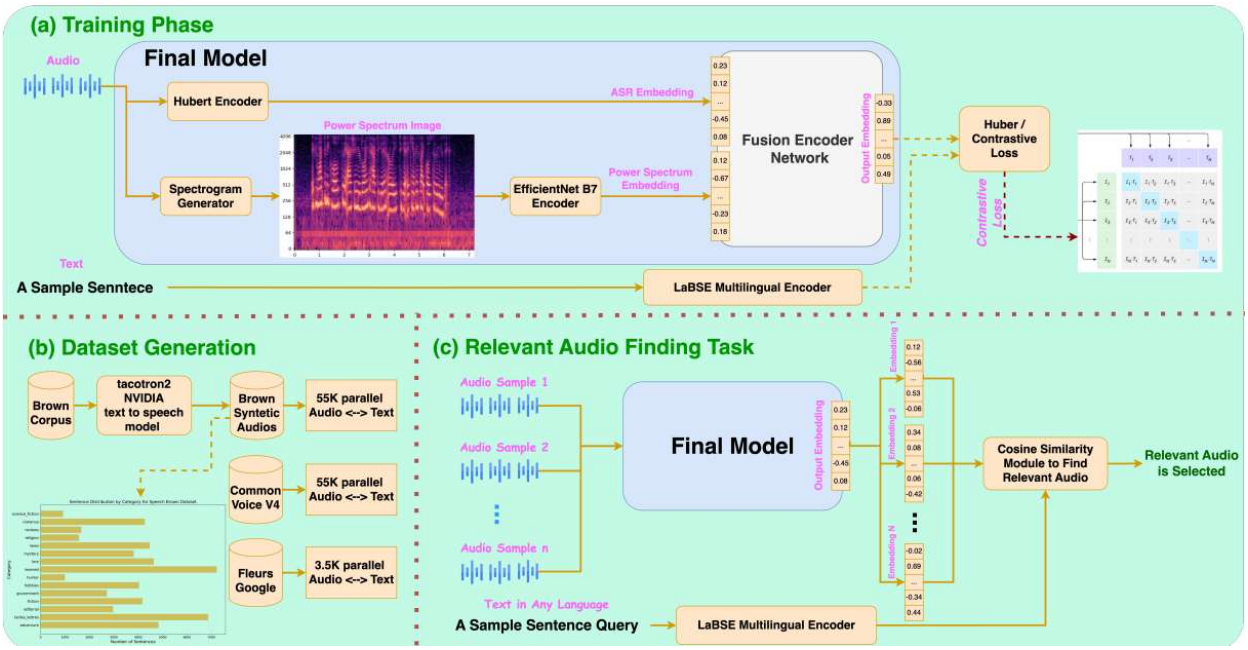
Title of the Paper

CLASP: Contrastive Language-Speech Pretraining for Multilingual Multimodal Information Retrieval

Summary of the Paper

This paper introduces **CLASP (Contrastive Language-Speech Pretraining)**, a multilingual and multimodal model designed for audio-text information retrieval. The authors propose a unified framework that integrates audio spectrograms, self-supervised speech encodings (HuBERT/Wav2Vec2), and pre-trained multilingual text encoders (XLM-R, LaBSE) to create a joint embedding space for speech and text. A key contribution is the introduction of **Speech Brown**, a diverse synthetic speech-text dataset spanning 15 domains, which enhances the model's linguistic diversity and robustness. CLASP is trained with contrastive loss, enabling it to learn from both positive and negative samples, resulting in more discriminative embeddings. The model outperforms traditional ASR-based systems in retrieval accuracy (HITS@1, MRR, meanR) and demonstrates strong multilingual capabilities across nearly 100 languages. Additionally, CLASP is more lightweight and efficient, with a ~50% reduction in model size and ~10% improvement in inference speed compared to ASR-based pipelines. While the model shows promise, future work could explore its scalability to large or real-time datasets, integration of speaker attributes, and applications beyond information retrieval, such as emotion recognition and speech understanding.

Main Architecture (or Idea) of the Paper



The core idea of CLASP is to align speech and text embeddings in a shared multilingual space. This is achieved by combining audio spectrograms, self-supervised speech models (HuBERT/Wav2Vec2), and multilingual text encoders (XLM-R, LaBSE). The model leverages contrastive learning to enhance the alignment of embeddings, enabling efficient and accurate audio-text retrieval across multiple languages. The introduction of the Speech Brown dataset further enriches the training process, ensuring robustness and diversity in the model's performance. This architecture allows CLASP to bridge the gap between speech and text modalities, making it a powerful tool for multilingual information retrieval.

Technical Strengths

1. **Multimodal Architecture**: The integration of spectrogram and self-supervised speech embeddings significantly improves the semantic representation of audio. This fusion of

modalities allows the model to capture richer and more nuanced information from both speech and text.

2. **Contrastive Learning**: The use of contrastive loss effectively aligns speech and text embeddings, enabling the model to learn from both positive and negative samples. This approach outperforms regression-based methods, resulting in more discriminative and robust embeddings.
3. **Multilingual Generalization**: The model supports cross-language retrieval, validated across four non-English languages, thanks to the use of LaBSE and XLM-R encoders. This capability makes CLASP a versatile tool for multilingual applications.
4. **Lightweight and Efficient**: CLASP achieves comparable or better performance than ASR-based systems while reducing computational cost and model size by ~50% and improving inference speed by ~10%. This efficiency makes it a practical solution for real-world deployment.
5. **Dataset Contribution**: The introduction of the Speech Brown dataset enhances training diversity and robustness. This dataset, spanning 15 domains, is a valuable resource for the research community and contributes to the model's ability to generalize across different contexts.

Technical Weaknesses

1. **Limited Task Scope**: The model is primarily evaluated for information retrieval, leaving its performance in ASR or generation tasks unexplored. While CLASP excels in retrieval, its applicability to other speech-related tasks remains unclear.
2. **Real-world Scalability**: The study does not assess the model's performance on real-time or large-scale streaming data, which is critical for practical deployment. This limitation raises questions about CLASP's ability to handle real-world scenarios.
3. **No Fine-tuning for Non-English Audio**: While the model supports multilingual text queries, the speech encoders are not fine-tuned for specific languages. This lack of fine-tuning may limit the model's performance in non-English audio retrieval.
4. **No Ablation for Fusion Strategies**: The comparison between gating and concatenation fusion strategies is limited, and a deeper analysis of their effects is missing. A more thorough examination of these strategies could provide insights into optimizing the model's architecture.

Minor Questions/Minor Weaknesses

1. **Handling Noisy or Low-Quality Audio**: How does CLASP handle noisy or low-quality audio during inference? The paper does not address this issue, which is crucial for real-world applications where audio quality may vary.
2. **Domain-Specific Performance**: Could the model's performance degrade in domain-specific audio (e.g., medical or legal jargon) not covered by Speech Brown? The dataset's diversity is a strength, but its coverage of specialized domains is unclear.

3. **Generalization to Unseen Languages**: While the model supports nearly 100 languages, its performance on unseen or low-resource languages is not explored. This raises questions about its ability to generalize to languages not included in the training data.

Suggestions for Improvement

To enhance the model's utility and credibility, future work should investigate CLASP's performance in low-resource or real-time streaming settings. Introducing noise-augmented or domain-adapted training could improve generalization to real-world audio. Expanding the model to cover ASR or speech summarization would demonstrate its flexibility and broaden its applicability. Additionally, a more detailed analysis of fusion strategies (e.g., attention-based vs gating vs concatenation) would help optimize the architectural design and provide insights into the most effective approaches for embedding alignment.

Rating and Justification

Rating: 8.5/10 The paper offers a technically sound, efficient, and novel approach for multilingual audio-text retrieval, backed by rigorous evaluation and dataset contribution. It loses points slightly for the lack of real-world applicability tests and limited task generalization. However, its strengths in multimodal integration, contrastive learning, and multilingual capabilities make it a significant contribution to the field.

https://arxiv.org/abs/2412.13071