Proceedings of NPUA: Information technologies, Electronics, Radio engineering. 2025. № 1.
UDC 004.832
DOI: 10.53297/18293336-2025.1-40

SEMANTIC CLUSTERING AND MULTI-MODEL INTEGRATION FOR EFFICIENT AUDIO CAPTIONING

E.A. Harutyunyan

National Polytechnic University of Armenia
Krisp

Audio captioning models play a crucial role in bridging the gap between acoustic information and human language, making digital audio content more accessible and searchable for diverse applications. These systems are increasingly vital in assistive technologies, content management, and surveillance systems where automated understanding of soundscapes is required. This research has introduced an innovative hybrid captioning methodology that boosts the capabilities of resource-efficient models without substantially increasing their computational footprint. The proposed approach harnesses the advantages of two lightweight audio captioning systems (Whisper-small and CoNeTTE) through a sophisticated pipeline encompassing multiple stages: initial caption generation, semantic phrase extraction, clustering of related concepts, selection of optimal phrases, and coherent text assembly. This technique enables the creation of richer, more detailed captions by combining the strongest elements from each model's output. Testing on the Clotho dataset revealed significant performance improvements, with the hybrid system surpassing individual models by substantial margins across all evaluation metrics. On average, the hybrid approach demonstrated enhancements of 28,4% over Whisper-small and 34,3% over CoNeTTE. Particularly impressive gains were observed in METEOR (48,4%) and SPICE (40,4%) metrics, highlighting the hybrid system's superior semantic accuracy and alignment with human-generated descriptions. These findings support our initial hypothesis that different architectures capture complementary aspects of audio content, with Whisper-small excelling in precision and CoNeTTE in semantic comprehension. Future research directions include expanding the framework with additional specialized models while refining the semantic clustering with adaptive thresholds.

Keywords: audio captioning, large language models, Whisper-small, CoNeTTE.

Introduction. Audio captioning is the process of automatically generating textual descriptions from audio recordings. This emerging field combines audio processing with natural language generation to create human-readable descriptions of sounds, music, and acoustic events (Fig. 1). Audio captioning systems have gained significant attention due to their ability to make audio content accessible and searchable [1]. These systems serve as valuable applications in media accessibility for individuals with hearing impairments, content indexing for audio databases, and automated surveillance of acoustic environments.

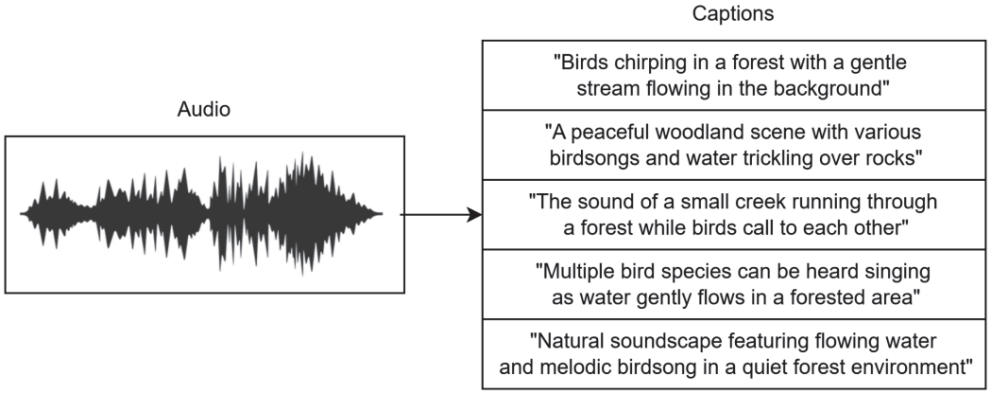

Fig. 1. Human caption variation for audio data

Despite its importance, audio captioning faces several challenges, particularly when deploying models with limited computational resources. Small models often struggle with accurately recognizing complex sound events, capturing temporal relationships, and generating natural-sounding descriptions. The field has recently witnessed a shift toward Large Language Models (LLMs) that demonstrate impressive captioning capabilities but require substantial computational resources [2]. While large-scale audio-language models achieve state-of-the-art performance, they demand high-end GPUs with significant memory capacity, making them impractical for many real-world applications.

Small-footprint models offer advantages in deployment flexibility, reduced energy consumption, and accessibility across diverse hardware configurations. However, these lightweight models generally produce less accurate captions compared to their larger counterparts. This research explores a novel approach to enhance the performance of small audio captioning models without significantly increasing their computational requirements through an intelligent hybrid pipeline.

Literature Review. Recent advancements in audio captioning have explored various innovative approaches to overcome limitations of lightweight models. Researchers have particularly focused on leveraging larger language models, retrieval techniques, and novel training paradigms to enhance caption quality without excessive computational requirements.

In [3], enhancing audio captioning is proposed by extensively leveraging pretrained models and LLMs. They utilized the BEATs Transformer for extracting fine-grained audio features and employed INSTRUCTOR LLM to generate text embeddings, incorporating this language knowledge through an auxiliary InfoNCE loss. Their innovative data augmentation technique used ChatGPT to create "caption mix-ups" paired with corresponding audio mixtures to increase the training data diversity. During inference, they implemented nucleus sampling and a hybrid reranking algorithm to select optimal captions from multiple candidates.

[4] suggested a solution to address the limited availability of audio-text

paired data by leveraging frozen pre-trained language models for audio captioning. Rather than training the entire model from scratch, they kept the language model frozen to preserve its text generation capabilities while only training components that extract audio features. Their key innovation was introducing mapping networks that translate audio features into "prefixes" - continuous vectors that effectively bridge the modality gap between audio and text representations. This approach prevented overfitting to small-scale datasets by utilizing the rich language understanding already embedded in pre-trained models.

In [5], RECAP (REtrieval-Augmented Audio CAPtioning) is implemented a novel system that generates captions conditioned by both input audio and similar captions retrieved from a replaceable datastore. Their approach leverages the CLAP audio-text model to retrieve captions similar to the input audio, which are then used to construct prompts fed into a GPT-2 decoder. A key advantage of RECAP is its ability to transfer to new domains without additional fine-tuning, while also demonstrating unique capabilities in captioning novel audio events never encountered during training. The method shows particular strength in handling compositional audios containing multiple events by exploiting a text-captions-only datastore.

Research methodology. In this study, an effective captioning pipeline is proposed specifically designed to improve textual outputs generated by small-scale audio captioning models. The Clotho dataset [6], a widely recognized audio captioning dataset comprising thousands of diverse audio clips, each paired with multiple natural language descriptions, was primarily utilized. This dataset is particularly valuable due to its rich variety of sounds, including environmental, human-made, animal, and indoor sounds. Such variety helps evaluate how effectively the proposed approach generalizes across different sound categories.

In this study, a hybrid captioning method is proposed to enhance textual outputs generated by small-scale audio captioning models. The approach utilizes outputs from multiple captioning models, combining them into a single, comprehensive caption (Fig. 2).

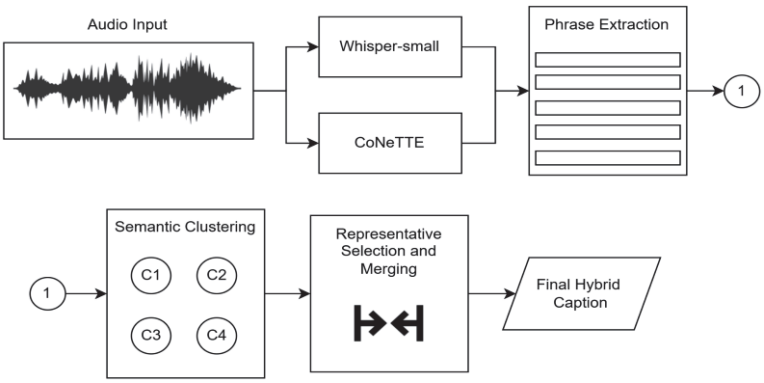

Fig. 2. Hybrid audio captioning pipeline diagram

Caption Generation and Phrase Extraction. Initially, each audio input is processed separately by three lightweight captioning models: Whisper-small [7], CoNeTTE [8], and a third lightweight model. Each model independently generates its own textual caption describing the given audio input. These captions are then cleaned, tokenized, and divided into meaningful semantic phrases. Phrase extraction involves breaking down each generated caption into smaller, coherent parts such as nouns, verbs, or verb phrases, which describe specific audio events or characteristics.

Semantic Clustering. After extracting individual phrases from each model's captions, semantic clustering is performed. This clustering groups similar or related phrases from different model outputs based on their semantic similarity. Phrase embeddings, which capture semantic meaning numerically, are computed using methods such as Sentence-BERT. Phrases with embeddings that show high semantic similarity (e.g., cosine similarity greater than 0.7) are grouped into clusters, ensuring that each cluster represents a coherent and distinct semantic event or concept identified by multiple models.

Representative Phrase Selection. From each semantic cluster, a representative phrase is selected to be included in the final hybrid caption. This selection aims to choose the most informative, accurate, and fluent phrase from each cluster. The criteria for choosing the representative phrase can include factors such as semantic completeness, grammatical correctness, clarity, and overall descriptive richness.

Fluent Merging. The selected representative phrases from each cluster are then merged into a single, cohesive caption. This final merging step focuses on ensuring that the hybrid caption is fluent, logically structured, and easily readable. The merging process may involve adding conjunctions or transition words, reordering phrases for better coherence, and minor grammatical adjustments to produce a caption that effectively integrates the strengths of all original model outputs.

Results. Experiments were conducted using the test portion of the Clotho dataset, containing diverse audio clips with multiple ground-truth human-generated descriptions per audio file. Common captioning evaluation metrics such as BLEU-4, METEOR, CIDEr, ROUGE, and SPICE [9] were computed to assess the caption quality quantitatively.

- **CIDEr**: Measures the consensus of captions based on the TF-IDF weighting of n-grams:

$$CIDEr - D(c,S) = \sum_{n=1}^{N} \quad w_n \cdot \frac{1}{m} \sum_{j=1}^{m} \quad \frac{gmn(c_i,sij)}{gmn(c_i,sij)+gmn(sij,c_i)}; \qquad (1)$$

- **BLEU-4**: A metric measuring precision of overlapping 4-grams between the candidate caption and references:

$$BLEU - 4 = BP \cdot exp\ exp\left(\sum_{n=1}^{4} w_n\ log\ log\ p_n\right); \qquad (2)$$

- **ROUGE-L**: Focuses on recall, evaluating how well the candidate caption captures the essential phrases in reference captions:

$$ROUGE - L = \frac{(1 + \beta^2) \cdot R_{lcs} \cdot P_{lcs}}{R_{lcs} + \beta^2 \cdot P_{lcs}}; \quad (3)$$

- **SPICE**: Evaluates the semantic propositional content similarity between the generated captions and references:

$$SPICE = \frac{2 \cdot |T(c) \cap T(s)|}{|T(c)| + |T(s)|}; \quad (4)$$

- **METEOR**: Evaluates captions based on alignment matching between candidate and references, accounting for synonyms and stemming:

$$METEOR = F_{mean} \cdot (1 - Penalty). \quad (5)$$

The evaluation involved comparing captions generated independently by Whisper-small and CoNeTTE, along with the hybrid method proposed by us. Each generated caption was evaluated against all five human-annotated reference captions per audio clip, with the scores representing averaged results across the entire test set (Table).

Table

Audio Captioning Model Performance Comparison

Model	**BLEU-4**	**METEOR**	**CIDEr**	**ROUGE-L**	**SPICE**
Whisper-small	0,221	0,256	0,537	0,387	0,161
CoNeTTE	0,202	0,211	0,526	0,402	0,175
Hybrid Method	0,269	0,380	0,590	0,470	0,226

The results demonstrated clear advantages of the hybrid captioning method across all metrics, with an average improvement of 28,4% over Whisper-small and 34,3% over CoNeTTE individually. Interestingly, the individual models exhibited different strengths across evaluation metrics. Whisper-small generally performed better on precision-oriented metrics (BLEU-4, METEOR, and CIDEr), suggesting superior accuracy in n-gram matching and word choice. Meanwhile, CoNeTTE demonstrated advantages in semantic and recall-oriented metrics (ROUGE-L and SPICE), indicating better performance in capturing overall meaning and sentence structure. The hybrid method successfully integrated these complementary

strengths, resulting in captions that excel in both precision and semantic richness.

Conclusion. This study presented a hybrid audio captioning method designed to enhance the performance of small-scale models without significantly increasing the computational demands. The approach leverages the complementary strengths of multiple lightweight audio captioning models (Whisper-small and CoNeTTE) through a novel pipeline that includes caption generation, phrase extraction, semantic clustering, representative phrase selection, and fluent merging. By extracting and combining the most informative phrases from each model's outputs, our approach creates more comprehensive and accurate captions that capture diverse aspects of audio content. Experimental evaluation on the Clotho dataset demonstrated that the hybrid method consistently outperformed individual models across all standard metrics, with an average improvement of 28,4% over Whisper-small and 34,3% over CoNeTTE. Performance gains were particularly notable in METEOR (48,4% improvement) and SPICE (40,4% improvement), indicating enhanced semantic richness and alignment with human descriptions. The results confirmed our hypothesis that different models excel in capturing different aspects of audio content, with Whisper-small showing strengths in precision-oriented metrics and CoNeTTE in semantic and recall-oriented metrics. Future work could explore extending the hybrid approach to include additional lightweight models, potentially incorporating domain-specific captioning experts for different audio categories. Investigating adaptive semantic clustering thresholds based on audio complexity and developing more sophisticated phrase selection criteria could yield further improvements. The methodology could also be extended to other multimodal tasks where combining outputs from specialist models might enhance overall performance while maintaining computational efficiency.

References

1. Training audio captioning models without audio / **S. Deshmukh, B. Elizalde, D. Emmanouilidou, B. Raj, et al** // In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). - 2024, April. - P. 371-375.
2. CoLLM: A Collaborative LLM Inference Framework for Resource-Constrained Devices / **J. Li, B. Han, S. Li, et al** // In 2024 IEEE/CIC International Conference on Communications in China (ICCC). - 2024, August. - P. 185-190.
3. Improving audio captioning models with fine-grained audio features, text embedding supervision, and llm mix-up augmentation / **S. Wu, X. Chang, G. Wichern, J. Jung, et al** // In ICASSP 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). - 2024, April. - P. 316-320.
4. **Kim M., Sung-Bin K., Oh T.H.** Prefix tuning for automated audio captioning // In ICASSP 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). - 2023, June. - P. 1-5.
5. Recap: Retrieval-augmented audio captioning / **S. Ghosh, S. Kumar, C. Evuru, et al** // In ICASSP 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). - 2024, April. - P. 1161-1165.

6. **Drossos K., Lipping S., Virtanen T.** Clotho: An audio captioning dataset // In 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). - 2020. - P. 736-740.
7. **Kadlčík M., Hájek A., Kieslich J., Winiecki R.** A whisper transformer for audio captioning trained with synthetic captions and transfer learning: arXiv preprint arXiv:2305.09690. - 2023.
8. **Pellegrini T., Pinquier J**. CoNeTTE: An efficient Audio Captioning system leveraging multiple datasets with Task Embedding. - 2023. - P. 1-10.
9. **Jaiswal S., Pallthadka H., Chinchewadi R., Jaiswal T**. An Extensive Analysis of Image Captioning Models, Evaluation Measures, and Datasets // International Journal of Multidisciplinary Science Research Review. - 2023. - P. 21-37.

Received on 23.05.2025.
Accepted for publication on 10.07.2025.

ԻՄԱՍՏԱՅԻՆ ԿԼԱՍՏԵՐԱՎՈՐՈՒՄ ԵՎ ԲԱԶՄԱՄՈԴԵԼԱՅԻՆ ԻՆՏԵԳՐՈՒՄ ԱՐԴՅՈՒՆԱՎԵՏ ՁԱՅՆԱՅԻՆ ՆԿԱՐԱԳՐՄԱՆ ՀԱՄԱՐ

Է.Ա. Հարությունյան

Ձայնային նկարագրման մոդելները կարևոր դեր են խաղում ակուստիկ տեղեկատվության և մարդկային լեզվի միջև կապ ստեղծելու գործում՝ թվային ձայնային բովանդակությունը դարձնելով ավելի մատչելի և որոնելի տարբեր կիրառությունների համար: Այս համակարգերն ավելի ու ավելի կարևոր են դառնում օժանդակ տեխնոլոգիաներում, բովանդակության կառավարման և հսկողության համակարգերում, որտեղ պահանջվում է ձայնային ազդանշանների ավտոմատացված ընկալում: Այս հետազոտությունը ներկայացնում է հիբրիդային նկարագրման մեթոդաբանություն, որը բարձրացնում է ռեսուրսային արդյունավետ մոդելների հնարավորությունները՝ առանց էականորեն ավելացնելու նրանց հաշվողական բարդությունը: Առաջարկվող մոտեցումն օգտագործում է երկու թեթև ձայնային նկարագրման միջոցների (Whisper-small և CoNeTTE) առավելությունները՝ բազմաթիվ փուլեր ներառող բարդ պրոցեսի միջոցով. սկզբնական նկարագրի ստեղծում, իմաստաբանական արտահայտությունների դուրսբերում, հարակից հասկացությունների կլաստերացում, օպտիմալ արտահայտությունների ընտրություն և կապակցված տեքստի հավաքագրում: Այս տեխնիկան հնարավորություն է տվել՝ ստեղծելու ավելի հարուստ, ավելի մանրամասն նկարագրություններ՝ համակցելով յուրաքանչյուր մոդելի արդյունքի ուժեղ տարրերը: Clotho տվյալների հավաքածուի վրա փորձարկումը ցույց է տվել կատարողականության զգալի բարելավումներ, ընդ որում հիբրիդային համակարգը գերազանցել է առանձին մոդելները բոլոր գնահատման ցուցանիշներով: Հիբրիդային մոտեցումը ցույցադրել է միջինը 28,4% բարելավումներ՝ Whisper-small-ի և 34,3% CoNeTTE-ի համեմատ: Առանձնապես տպավորիչ աճ է նկատվել METEOR (48,4%) և SPICE (40,4%) ցուցանիշներում, ընդգծելով հիբրիդային միջոցի բարելավված իմաստաբանական ճշտությունը և մարդու կողմից ստեղծված նկարագրությունների հետ համապատասխանությունը: Այս արդյունքները հաստատում են նախնական վարկածը, որ տարբեր ճարտարապետություններ ընդգրկում են ձայնային բովանդակության լրացուցիչ ասպեկտներ, ընդ որում, Whisper-small-ը գերազանցում է ճշտության, իսկ CoNeTTE-ն՝ իմաստաբանական ըմբռնման մեջ: Ապագա հետազոտությունների

ուղղությունները ներառում են այս համակարգի ընդլայնումը լրացուցիչ մասնագիտացված մոդելներով, միաժամանակ կատարելագործելով իմաստաբանական կլաստերացումը ադապտիվ շեմերի միջոցով:

Առանցքային բառեր. ձայնային նկարագրություն, մեծ լեզվական մոդելներ, Whisper-small, CoNeTTE:

СЕМАНТИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ И МУЛЬТИМОДЕЛЬНАЯ ИНТЕГРАЦИЯ ДЛЯ ЭФФЕКТИВНОГО ЗВУКОВОГО ОПИСАНИЯ

Э.А. Арутюнян

Модели звукового описания играют ключевую роль в преодолении разрыва между акустической информацией и человеческим языком, делая цифровой звуковой контент более доступным и легко находимым для различных приложений. Эти системы становятся всё более важными во вспомогательных технологиях, управлении контентом и системах наблюдения, где требуется автоматизированное понимание звуковых ландшафтов. В данном исследовании представлена инновационная методология гибридного описания, которая повышает возможности ресурсоэффективных моделей без существенного увеличения их вычислительной нагрузки. Предложенный подход использует преимущества двух легковесных систем звукового описания (Whisper-small и CoNeTTE) через сложный конвейер, охватывающий несколько этапов: начальную генерацию описания, извлечение семантических фраз, кластеризацию связанных концепций, выбор оптимальных фраз и связное составление текста. Эта техника позволяет создавать более богатые и детальные описания путем интеллектуального объединения сильных элементов из выходных данных каждой модели. Тестирование на базе данных Clotho выявило значительные улучшения производительности, при этом гибридная система превосходила отдельные модели по всем метрикам оценки. В среднем гибридный подход продемонстрировал улучшения на 28,4% по сравнению с Whisper-small и на 34,3% по сравнению с CoNeTTE. Особенно впечатляющие успехи наблюдались в метриках METEOR (48,4%) и SPICE (40,4%), подчеркивая превосходную семантическую точность гибридной системы и соответствие описаниям, созданным человеком. Эти результаты подтверждают нашу первоначальную гипотезу о том, что различные архитектуры охватывают дополняющие аспекты звукового контента, при этом Whisper-small превосходит в точности, а CoNeTTE - в семантическом понимании. Направления будущих исследований включают расширение фреймворка с дополнительными специализированными моделями, а также совершенствование семантической кластеризации с помощью адаптивных порогов.

Ключевые слова: звуковое описание, большие языковые модели, Whisper-small, CoNeTTE.