MULTIMODAL HATR CLASSIFICATION WITH FROZEN CLAP EMBEDDINGS FOR THE BROAD SOUND TAXONOMY

Technical Report

Han Zheng and Yanxiong Li

School of Electronic and Information Engineering
South China University of Technology
Guangzhou, China
13996733139@163.com; eeyxli@scut.edu.cn

ABSTRACT

This report presents a multimodal audio classifier for DCASE 2026 Task 1, which assigns Freesound recordings to 23 second-level categories of the Broad Sound Taxonomy (BST). The system combines frozen LAION-CLAP embeddings of audio and textual metadata (title, tags, description) with a Hierarchical Audio Tagging and Retrieval (HATR) feed-forward classifier using attention-based fusion. Development performance is measured under five-fold cross-validation on BSD10k-v1.2 and BSD35k-CS ($\lambda = 0.75$). On BSD10k-v1.2, the multimodal model attains macro-averaged hierarchical F-scores of $78.65\% \pm 0.52$ (hF), compared with $75.88\% \pm 0.47$ for the audio-only configuration; on BSD35k-CS, the corresponding values are $79.77\% \pm 0.60$ and $70.19\% \pm 0.93$, with leaf accuracy reaching $85.32\% \pm 0.51$. For the challenge evaluation set, five fold models trained on BSD10k-v1.2 are ensembled to produce predictions for all 3246 released test recordings.

Index Terms— Broad Sound Taxonomy, contrastive language–audio pretraining, heterogeneous audio classification, hierarchical evaluation, multimodal fusion

1. INTRODUCTION

Environmental sound classification maps acoustic recordings to semantic categories and underpins a wide range of practical systems. Deep models have been developed for label-scarce and continually evolving category sets, including few-shot class-incremental learning with dynamically expanded decision boundaries [1] and adaptive mitigation of forgetting under incremental updates [2, 3]. Open-set and pseudo-incremental embedding strategies further extend this line of work [4, 5].

Representative applications include multimedia content analysis [6], automated audio captioning [7], and acoustic monitoring of transportation infrastructure [8]. Acoustic scene analysis has also been applied to home security [9], often in combination with sound event detection backbones [10].

The DCASE 2026 heterogeneous audio classification task requires assigning each recording to one of 23 second-level categories defined by the Broad Sound Taxonomy (BST) [11]. Unlike fine-grained ontologies such as AudioSet, the BST organizes everyday and professionally recorded sounds into five top-level groups—music, human sounds, natural sounds, soundscapes, and sound effects—each subdivided into a small set of interpretable leaf classes. This design supports practical retrieval and browsing applications, but it also poses a classification problem in which class boundaries are broad, recordings are highly variable in duration and signal quality, and semantic overlap between categories is non-negligible.

Challenge organizers evaluate submissions with macro-averaged hierarchical precision, recall, and F-score (hP, hR, hF) [12]. The weighted variant implemented in the baseline code assigns partial credit when a prediction is wrong at the leaf level yet correct at the top level, controlled by parameter λ. With $\lambda = 0.75$, confusions within the same top-level branch are penalized less severely than confusions across branches, aligning the metric with the two-level taxonomy structure.

The present system follows the publicly released multimodal baseline for DCASE 2026 Task 1 [13], adopting its CLAP-based feature extraction and five-fold training protocol. The HATR classifier design [14] serves as a reproducible starting point. Section 3–7 detail the feature extraction stage, classifier architecture, development results on both official corpora, and evaluation-set inference pipeline.

2. TASK FORMULATION AND DATA

Each training instance consists of a monophonic waveform sampled at 44.1 kHz (24-bit, maximum duration 30 s), together with textual metadata fields *title*, *tags*, and *description*. The learning target is the second-level BST label $y \in \mathcal{C}$, with $|\mathcal{C}| = 23$. Two development corpora are considered in this report.

BSD10k-v1.2 comprises roughly 11 000 expert-annotated recordings with imbalanced class priors and optional confidence scores. Annotations were curated for taxonomy consistency and serve as the training source for the submitted evaluation model.

BSD35k-CS contains approximately 35 000 crowd-sourced labels assigned by Freesound uploaders. Labels are noisier and the class distribution differs from BSD10k-v1.2, but the corpus offers greater coverage of in-the-wild naming variability. It is evaluated here under the same classifier architecture to characterize baseline behaviour on large-scale weak supervision, but is not used to train the submitted system.

Model selection and reporting use five-fold cross-validation with fixed random seed 1821 on each corpus independently. The released evaluation set contains 3246 audio files with anonymized identifiers and metadata, without public labels.

Corresponding author: Yanxiong Li

3. FEATURE EXTRACTION

3.1. Audio and text encoders

Both modalities are represented with the publicly released LAION-CLAP model `630k-audioset-fusion` [15]. CLAP learns aligned representations of audio and text through large-scale contrastive pretraining; the fusion variant aggregates multiple temporal and spectral views before projection to a shared embedding space. In this submission the CLAP weights remain fixed: no gradient updates are applied to the pretrained encoders during BSD10k training or evaluation-set processing.

For each audio file, a 512-dimensional vector $\mathbf{a} \in \mathbb{R}^{512}$ is extracted from the waveform. For text, the three metadata fields are concatenated in field order and passed through CLAP's text tower, yielding $\mathbf{t} \in \mathbb{R}^{512}$. On the development set, precomputed embeddings distributed with BSD10k-v1.2 are used for training. For the evaluation set, embeddings are computed locally with the same checkpoint (`630k-audioset-fusion-best.pt`) and cached prior to classification.

3.2. Implementation notes

Audio shorter than the CLAP analysis window is processed without manual padding beyond the model's internal handling. During embedding extraction, batch failures were observed for extremely short clips; those items were reprocessed individually to avoid discarding valid test samples. One evaluation waveform file was empty (0 bytes); for that identifier the text embedding was retained and the audio vector was set to zero, allowing the classifier to fall back on metadata alone.

4. CLASSIFIER ARCHITECTURE

The trainable model is a HATR-style feed-forward classifier operating on CLAP embeddings [14]. It follows the Task 1 baseline implementation [13] and contains approximately 7.32 million parameters and does not introduce explicit hierarchical decision layers: a single softmax head predicts leaf labels directly.

4.1. Modality-specific encoding

Each embedding passes through an `EmbeddingEncoder` composed of a linear projection, three residual blocks with LeakyReLU activations, batch normalization, and dropout ($p = 0.1$), followed by a bottleneck projection to hidden dimension $d = 128$:

$$\mathbf{h}^{(m)} = f_{\theta_m}(\mathbf{e}^{(m)}), \quad m \in \{\text{audio}, \text{text}\}. \tag{1}$$

4.2. Attention-based fusion

In `both` mode, encoded audio and text vectors are combined by an attention fusion module. The concatenated representation $[\mathbf{h}^{(\text{audio})}; \mathbf{h}^{(\text{text})}]$ is mapped to two nonnegative weights α_a, α_t that sum to one, and the fused feature is their weighted sum:

$$\mathbf{h}^{(\text{fused})} = \alpha_a \mathbf{h}^{(\text{audio})} + \alpha_t \mathbf{h}^{(\text{text})}. \tag{2}$$

A latent projector and two residual blocks further transform $\mathbf{h}^{(\text{fused})}$ before a linear layer outputs logits $\mathbf{z} \in \mathbb{R}^{23}$. Training optimizes cross-entropy against leaf labels; misclassifications that remain within the correct top-level branch are not explicitly rewarded at training time.

5. TRAINING AND MODEL SELECTION

All reported development experiments use the `both` modality unless stated otherwise. Optimization employs Adam with learning rate 10^{-3} and mini-batches of size 64. Training runs for up to 100 epochs with early stopping on validation hierarchical F1; the checkpoint maximizing that criterion on the held-out fold partition is retained. No data augmentation, mixup, or label smoothing is applied.

The **submitted evaluation model** is trained exclusively on BSD10k-v1.2. Five fold-specific checkpoints are retained and their class probabilities averaged at inference:

$$\hat{p}(y = c \mid \mathbf{x}) = \frac{1}{5} \sum_{k=1}^{5} p_k(y = c \mid \mathbf{x}). \tag{3}$$

The argmax of $\hat{p}$ defines the submitted label; the reported score is the corresponding maximum probability. Identical hyperparameters and architecture are applied when re-running the baseline on BSD35k-CS for comparative development-set analysis.

6. DEVELOPMENT-SET RESULTS

Table 1 summarizes five-fold mean performance on both development corpora ($\lambda = 0.75$). On BSD10k-v1.2, incorporating text metadata improves hF by 2.77 points over the audio-only variant ($78.65\% \pm 0.52$ vs. $75.88\% \pm 0.47$). On BSD35k-CS, the multimodal system achieves $79.77\% \pm 0.60$ hF and $85.32\% \pm 0.51$ leaf accuracy, outperforming audio-only training by 9.58 hF points. Notably, BSD35k-CS yields higher flat accuracy yet only a modest hF gain over BSD10k-v1.2, suggesting that crowd labels inflate leaf-match rates while hierarchical errors persist under noisy supervision.

6.1. BSD10k-v1.2

Per-fold hF values on BSD10k-v1.2 (both mode) are 78.81, 79.35, 78.05, 78.05, and 79.01 %. Table 2 compares class-wise hF across corpora. On BSD10k-v1.2, strong performance is observed for `is-w` (97.2 %), `is-s` (94.5 %), and `fx-o` (90.1 %); weak performance persists for `sp-c` (49.3 %), `ss-i` (55.4 %), and `fx-a` (63.6 %).

6.2. BSD35k-CS

Per-fold hF values on BSD35k-CS (both mode) are 80.18, 79.57, 80.71, 79.15, and 79.22 %. Several classes improve markedly relative to BSD10k-v1.2—for instance `fx-a` (+15.5 pp), `ss-n` (+19.5 pp), and `m-m` (+17.5 pp)—while others degrade sharply, including `is-k` (49.8 %), `is-w` (64.9 %), and `sp-c` (36.7 %). This divergence is consistent with label noise and shifting class priors in crowd-sourced data rather than a uniform gain from dataset scale.

7. EVALUATION-SET PROCEDURE

Evaluation-set inference mirrors development-set decoding except that labels are unavailable and embeddings must be computed from the released audio and metadata. The output file `Zheng_SCUT_task1_1.output.csv` contains 3246 lines with fields `id`, `predicted_bst_second_level_class`, and `prediction_score`. Identifiers match the `anonymous_id`

Table 1: Five-fold mean development performance on BSD10k-v1.2 and BSD35k-CS.

	BSD10k-v1.2		BSD35k-CS	
Metric	both	audio	both	audio
hP	79.51 ± 0.69	77.23 ± 0.46	81.07 ± 1.69	72.38 ± 2.14
hR	78.32 ± 0.56	75.33 ± 0.56	79.41 ± 1.01	69.10 ± 0.44
hF	$\mathbf{78.65 \pm 0.52}$	75.88 ± 0.47	$\mathbf{79.77 \pm 0.60}$	70.19 ± 0.93
Leaf accuracy	79.98 ± 0.56	77.28 ± 0.43	85.32 ± 0.51	75.36 ± 0.21
Top-level accuracy	89.18 ± 0.24	87.83 ± 0.38	92.71 ± 0.42	87.18 ± 0.40

Table 2: Class-wise hierarchical F-score (%) on BSD10k-v1.2 and BSD35k-CS (5-fold mean, both mode).

Class	10k	35k	Class	10k	35k	Class	10k	35k	Class	10k	35k
fx-a	63.6	79.1	is-e	76.4	92.6	sp-c	49.3	36.7	ss-i	55.4	56.4
fx-el	74.9	87.9	is-k	93.9	49.8	sp-p	78.6	72.9	ss-n	72.1	91.6
fx-ex	60.5	82.8	is-p	89.0	94.7	sp-s	92.3	89.3	ss-s	80.6	79.3
fx-h	90.6	79.6	is-s	94.5	91.6	m-m	73.3	90.8	ss-u	70.3	85.8
fx-m	82.0	83.2	is-w	97.2	64.9	m-si	82.0	87.4			
fx-n	86.1	63.9	m-sp	87.6	89.6	fx-o	90.1	92.0	fx-v	68.9	92.8

column in the evaluation metadata. System description and development results are additionally encoded in `Zheng_SCUT_task1_1.meta.yaml`.

8. EXTERNAL RESOURCES AND COMPLIANCE

Training for the submitted system relies exclusively on BSD10k-v1.2. Development-set results on BSD35k-CS are included for baseline comparison only. The only external pretrained component is LAION-CLAP; no Freesound recordings uploaded after 1 April 2025 enter the training pipeline.

9. DISCUSSION

Cross-corpus comparison highlights complementary failure modes. BSD10k-v1.2 favours expert-validated categories such as winds and sustained instruments, whereas BSD35k-CS improves several effect and speech-related classes while hurting others that likely suffer from inconsistent uploader labelling. The submitted configuration keeps CLAP frozen and therefore cannot correct systematic embedding mismatches; the gap between top-level accuracy (89.2 % on BSD10k-v1.2) and leaf hF further indicates that the flat classifier head is not aligned with the hierarchical evaluation measure. Reported numbers establish a reproducible reference on both official development corpora; subsequent work will examine embedding adaptation and explicit hierarchy modelling beyond this baseline.

10. REFERENCES

[1] Y. Li, W. Cao, W. Xie, J. Li, and E. Benetos, "Few-shot class-incremental audio classification using dynamically expanded classifier with self-attention modified prototypes," *IEEE Transactions on Multimedia*, vol. 26, pp. 1346–1360, 2024.

[2] Y. Li, J. Li, Y. Si, J. Tan, and Q. He, "Few-shot class-incremental audio classification with adaptive mitigation of forgetting and overfitting," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2297–2311, 2024.

[3] Y. Si, Y. Li, J. Tan, G. Chen, Q. Li, and M. Russo, "Fully few-shot class-incremental audio classification with adaptive improvement of stability and plasticity," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 418–433, 2025.

[4] Y. Li, W. Cao, J. Tan, Q. Li, and G. Chen, "Few-shot class-incremental audio classification using pseudo-incrementally trained embedding learner and continually updated stochastic classifier," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 3880–3895, 2025.

[5] Y. Li, J. Tan, Q. Li, G. Chen, S. Huang, and T. Virtanen, "Few-shot open-set audio classification using attention information-fused prototypes," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 1929–1943, 2026.

[6] W. Pang *et al.*, "Detecting video anomalies by jointly utilizing appearance and skeleton information," *Expert Systems with Applications*, vol. 246, p. 123135, 2024.

[7] Q. Li and Y. Li, "SCUT submission for automated audio captioning using graph attention and cross-attention mechanisms," DCASE 2024 Challenge Technical Report, 2024, https://dcase.community/documents/challenge2024/technical_reports/DCASE2024_Li_54_t6.pdf.

[8] Y. Li, X. Li, Y. Zhang, M. Liu, and W. Wang, "Anomalous sound detection using deep audio representation and a BLSTM network for audio surveillance of roads," *IEEE Access*, vol. 6, pp. 58 043–58 055, 2018.

[9] A. Chen, Q. He, X. Wang, and Y. Li, "Home security surveillance based on acoustic scenes analysis," in *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2017, pp. 1–5.

[10] Y. Li, M. Liu, K. Drossos, and T. Virtanen, "Sound event detection via dilated convolutional recurrent neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 286–290.

[11] P. Anastasopoulou, F. Font, D. Bogdanov, and L. Wyse, "Heterogeneous audio classification – DCASE 2026 task 1," https://dcase.community/challenge2026/task-heterogeneous-audio-classification, 2026.

[12] S. Kiritchenko, S. Matwin, and A. F. Famili, "Functional annotation of genes using hierarchical text categorization," in

Proceedings of the ACL Workshop on Linking Biological Literature, Ontologies and Databases, 2005.

[13] Music Technology Group, UPF, "DCASE2026 Task 1 Baseline System," https://github.com/MTG/dcase2026_task1_baseline, 2026.

[14] P. Anastasopoulou, F. Font, D. Bogdanov, and L. Wyse, "Hierarchical audio tagging and retrieval with the broad sound taxonomy," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2025.

[15] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.