HSLU Lucerne University of Applied Sciences and Arts

Technik und Architektur
Master Thesis

Identification of Sound Sources in Soundscapes using AI

HTS-AT fine-tuning for 10-second BAFU source detection

MSE Student: Lél Bartha

1. Background and research question

Soundscape quality depends on what is heard, not only on how loud it is. The BAFU Klangraum project investigates outdoor residential spaces with acoustic, room-acoustic, psychoacoustic, visual and social criteria.

This thesis evaluates whether HTS-AT, a hierarchical audio transformer, can detect multiple meaningful sound sources in long field recordings and whether it can localize them in time.

2. Material and methods

The initial Label Studio annotations were incomplete for event-level evaluation, so the dataset was relabeled before final model comparison. Evaluation was then split into three levels: full-recording clip ranking, new 10-second window clip ranking, and strict event-boundary metrics.

The new metric uses the exact 10-second windows sent to HTS-AT with 50% overlap. A class is positive in a window only if at least 3 seconds of ground-truth annotation overlap that window.

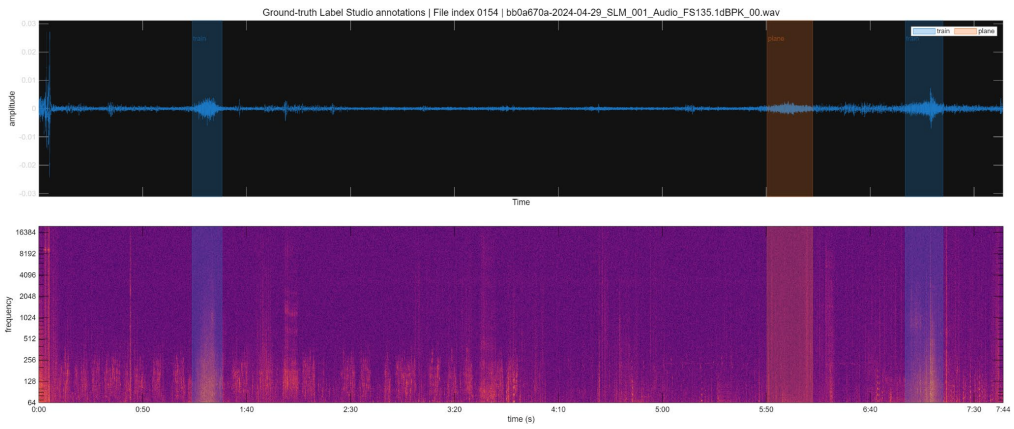

Figure 1: Old ground-truth overlay before relabeling; missing and background sources were corrected in the final labels.

Relabeling increased the cumulative annotated class duration from 10.55 h to 35.92 h and the number of labeled segments from 693 to 1,593. This was essential because missing labels made real audible sources appear as false positives during localization evaluation.

Several existing checkpoints for the HTS-AT model were used and fine tuned for the BAFU dataset. These checkpoints were either trained on the AudioSet dataset from Google, or the ESC-50 evaluation dataset. The best method of fine tuning was determined to be full weight and bias retraining, the results of which are in this poster.

4. Conclusion and recommendation

The thesis produced a functioning HTS-AT pipeline, a corrected multi-label dataset, and a comparison between file-level, 10-second window-level and event-boundary metrics.

The fully fine-tuned models are suitable for saying whether a source is present in a full recording and reasonably strong for detecting source presence in 10-second sections. They are not yet reliable for exact start/end localization. Further work should focus on class-specific calibration, more strong labels and thresholding after fine-tuning.

Industry partner: Bundesamt für Umwelt, BAFU, Dr. Trond Maag

3. Results and discussion

The final update changed the fine-tuning objective from event F1 to a 10-second window-level clip mAP. This keeps the threshold-independent AP/mAP formulation, but changes the question from "is this class somewhere in the whole recording?" to "in this particular 10 seconds of a long recording, is this sound source present for a meaningful amount of time?" measured from 0 to 1.

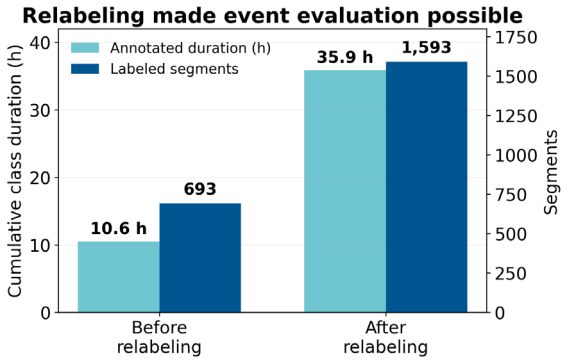

Figure 2: Relabeling greatly increased annotated duration and segment count.

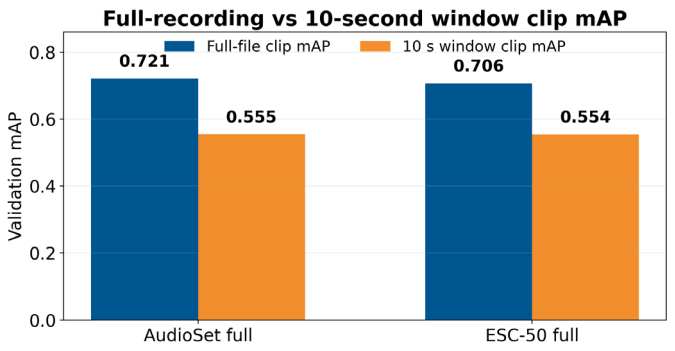

Figure 3: Full-recording clip mAP remains highest, but 10-second window mAP is the more relevant coarse temporal metric.

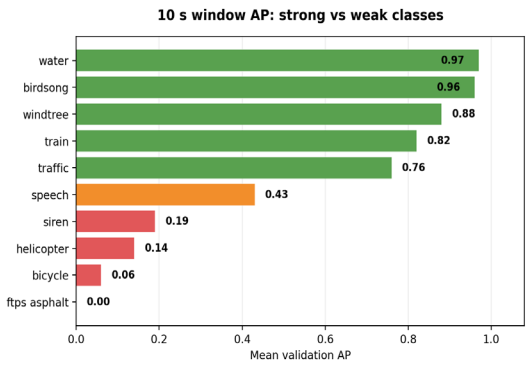

Figure 4: 10 s window AP highlights strong and weak classes.

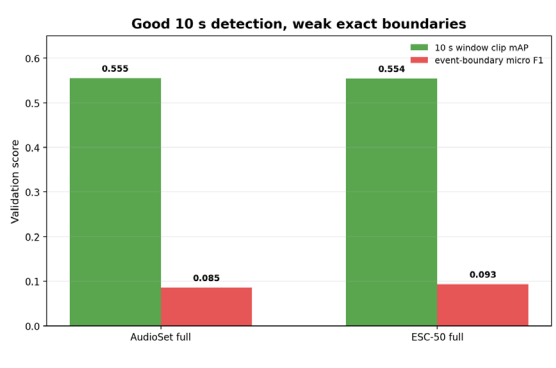

Figure 5: Coarse 10 s detection is far stronger than exact event-boundary localization.

Full fine-tuning is very strong for full-recording source presence: AudioSet full reached clip mAP = 0.721 and ESC-50 full reached 0.706. When the same models are evaluated on 10-second windows, performance drops to 0.555 and 0.554, which is expected because the model must now place the source in the right part of the long recording.

The 10-second results show useful coarse temporal source detection, especially for water, birdsong, windtree, train and traffic. Fine localization remains weak: event-boundary micro F1 is only 0.085 for AudioSet full and 0.093 for ESC-50 full, indicating the that detecting exact start and end times are not feasable with this setup.

Literature

Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T., & Dubnov, S. (2022). HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection.

HSLU / BAFU. (2025). Objektive Methoden der Klangraumbeurteilung und Möglichkeiten der KI-basierten Klangraumklassifizierung. Project report.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., et al. (2017). Audio Set: An ontology and human-labeled dataset for audio events. ICASSP.

Advisor: Prof. Dr. Armin Taghipour
Expert: Dr. Hervé Lissek
Semester: Spring 2026

Sources: BAFU Klangraum project · HTS-AT paper · thesis relabeling, fine-tuning and 10-second window mAP outputs