DCASE 2026 Task 4 Submission: Duplicate-Label-Aware Source Selection and TUSS Overlay Grafting

Yuhuan You
Peking University
2000017809@stu.pku.edu.cn

Abstract

This report describes the PKU submission to DCASE 2026 Challenge Task 4, Spatial Semantic Segmentation of Sound Scenes. The task requires detection and separation of target sound events from FOA mixtures, and the 2026 setting includes same-class multiple sources and zero-target soundscapes. Our submitted systems use the official baseline family as the separation and tagging foundation, then add duplicate-label-aware candidate selection and label-conditioned overlay grafting. The main ranking metric is CAPI-SDRi, so the system is optimized to select separated sources that jointly improve class assignment and permutation-invariant signal quality. We submit four systems: a stable duplicate-label weighted selector, two progressively more aggressive TUSS overlay systems, and a full all-label TUSS candidate-pool selector for diversity.

1 Task Setting

DCASE 2026 Task 4 evaluates spatial semantic segmentation of sound scenes. Each input is a 10 s first-order ambisonic mixture sampled at 32 kHz, and each output is a set of separated monaural event signals with predicted class labels. The 2026 task is harder than the previous version because a mixture can contain multiple sources from the same class. This affects both detection and separation because the label prediction stage must distinguish source multiplicity, and the separation stage is scored with a permutation-invariant CAPI-SDRi metric.

The evaluation package for each submitted system contains separated audio for the evaluation set, an `eval_results.json` file, and a `dev_set_test_results.json` file. The official CMT package contains metadata, this report, and a text file with links to the four output zip files. The submitted output zips follow the required names `You_PKU_task4_1_out.zip` through `You_PKU_task4_4_out.zip`.

2 Baseline Foundation

The base separation pipeline follows the DCASE Task 4 baseline design. M2D audio tagging features are used to estimate event labels, while ResUNet and ResUNetK variants provide separated source candidates. The official baseline already provides strong spatial audio handling through FOA inputs and a trained tagger-separator pairing. We use this family as a reliable foundation because it is aligned with the official data format, output convention, and development-set evaluation protocol.

The main weakness addressed by our system is the candidate selection problem. The same-class-source condition creates cases where a simple per-label or per-clip rule either misses duplicated events or keeps too many low-quality separated sources. We therefore treat each source hypothesis as a candidate with class, activity, duplicate-label, confidence, and signal-quality proxy features, and select candidates with calibration rules learned on the development test subset.

3 Duplicate-Label-Aware Selection

Our first submitted system is a duplicate-label-aware weighted selector. It combines multiple candidate streams from the baseline family and uses source-level metadata to decide whether a candidate should be emitted. The selector explicitly tracks whether a source belongs to an active class and whether the same class appears multiple times in the clip. These features matter because the CAPI-SDRi metric rewards correct source multiplicity and penalizes both missed target sources and mismatched extra sources.

The stable selector uses a weighted blend of two calibrated model families. The best development evidence for this line is a global CAPI-SDRi of 11.856, a grouped score of 11.859, and a nested grouped score of 11.853. The corresponding oracle score is 12.276, which shows that the candidate pool still contains additional upside, although the conservative selector already gives the most reliable full-development result among the tested systems.

4 TUSS Overlay Grafting

The second and third submitted systems add TUSS overlay grafting. TUSS is used as a query-conditioned external source separator, producing alternative candidate sources for classes where development-set ablations showed positive replacement behavior. Instead of replacing the whole system with TUSS, we graft selected TUSS sources into the stable baseline output. This keeps the robust duplicate-label selector as the backbone while allowing specific labels to use higher-quality external separations.

The second system applies all positive label overlays found during the development audit. The affected labels are Speech, Percussion, Cough, VacuumCleaner, Dishes, FootSteps, and MusicalKeyboard. On the evaluation output this corresponds to 185 source replacements, with the largest changes in Speech and Percussion. The third system starts from the second system and adds a small Blender-line overlay. This third system changes only six additional evaluation sources, so it is a low-magnitude risk line designed to keep most of the stronger all-positive overlay system unchanged while testing a small hidden-set improvement opportunity.

5 Full All-Label TUSS Candidate Selector

The fourth submitted system uses a larger all-label TUSS candidate pool. A global ExtraTrees-style selector with 1200 trees is calibrated on raw TUSS candidate features and baseline outputs. This system has a development CAPI-SDRi of 11.852, which is slightly below the stable duplicate-label weighted selector, but its oracle reaches 12.530. That higher oracle indicates that the all-label TUSS pool contains stronger separated sources than the conservative pool in some clips. We include it as a diversity submission because hidden evaluation may reward a different selector-candidate tradeoff.

6 Submitted Systems and Development Evidence

Label	Main idea	Dev CAPI-SDRi	Label accuracy
You_PKU_task4_1	Stable duplicate-label weighted selector	11.856	74.868
You_PKU_task4_2	Task 1 plus all-positive TUSS overlay grafting	11.856	74.868
You_PKU_task4_3	Task 2 plus a small Blender-line overlay	11.856	74.868
You_PKU_task4_4	Full all-label TUSS raw ET1200 selector	11.852	74.868

Table 1: Development evidence used for the final submission slate. The overlay systems inherit the stable selector evidence and are distinguished by local positive overlay audits and low-magnitude evaluation-set source replacements.

The four submitted systems are intentionally correlated around the best stable selector, while also preserving one larger candidate-pool diversity line. This slate is chosen because DCASE permits up to four systems and the final ranking is expected to be sensitive to hidden-set source distribution. The first system is the main conservative system. The second and third systems test whether development-positive TUSS overlays generalize to evaluation. The fourth system tests whether the wider all-label TUSS source pool can outperform the conservative pool under hidden-set conditions.

7 Conclusion

The submission focuses on the main technical change in DCASE 2026 Task 4: same-class sources and zero-target soundscapes. The system therefore emphasizes duplicate-label-aware source selection, calibrated candidate emission, and cautious source-level grafting from external TUSS candidates. The strongest full-development result is 11.856 CAPI-SDRi, while the all-label TUSS candidate pool shows higher oracle potential. The final four-system slate balances stable development performance with hidden-set diversity.

References

[1] DCASE 2026 Challenge Task 4, Spatial Semantic Segmentation of Sound Scenes. https://dcase.community/challenge2026/task-spatial-semantic-segmentation-of-sound-scenes

[2] NTT Communication Science Laboratories, DCASE 2026 Task 4 baseline. https://github.com/nttcslab/dcase2026_task4_baseline

[3] M2D self-supervised audio representation. https://github.com/nttcslab/m2d

[4] Unified Source Separation. https://github.com/merlresearch/unified-source-separation