Task 15 – Aggregation Implementation

Introduction

For this task, temporal averaging (mean aggregation) was used as the aggregation strategy. Each audio track was first segmented into overlapping fixed-length windows. The model produced segment-level probability predictions for each window. These probabilities were then averaged across all segments belonging to the same track to obtain a clip-level probability vector, which was finally thresholded to produce multi-label predictions.

This strategy was chosen because averaging over time is a widely used and interpretable aggregation method in audio tagging tasks. It reduces the influence of noisy or anomalous segment-level predictions while reinforcing patterns that persist across multiple segments.

Comparison: With Aggregation vs Without Aggregation

Quantitative Results

Strategy	Micro F1	Macro F1
Without Aggregation	0.5566	0.4758
With Aggregation	0.5686	0.4974

Both Micro F1 and Macro F1 improved when aggregation was applied.

Observed Effects of Aggregation

1. False Positives:

Without aggregation, predictions were made by voting over independently thresholded segments. This caused spurious activations, where a single noisy segment could trigger a false positive for an instrument that was not truly present throughout the clip.

With aggregation, such isolated spikes were averaged out. As a result:

- Precision improved for several classes (e.g., gac, pia, voi)
- Random false positives caused by short-lived noise or masking effects were reduced

This indicates that aggregation suppressed unstable segment-level activations.

2. Missed Detections (False Negatives):

Aggregation improved recall for several instruments, particularly those that:

- Appeared intermittently
- Were weak or partially masked in individual segments

By accumulating evidence over time, aggregation allowed the model to detect instruments even when they were not strongly present in every segment. This is reflected in the higher recall values for classes such as tru, vio, and gel.

Thus, aggregation reduced missed detections caused by short or fragmented instrument presence.

3. Stability of Predictions:

Without aggregation, segment-level predictions fluctuated significantly across time. Adjacent segments often produced contradictory outputs, leading to unstable clip-level decisions.

Aggregation produced:

- Smoother probability estimates
- More consistent clip-level predictions
- Reduced sensitivity to local noise or brief acoustic events

This temporal consistency aligns more closely with how humans perceive instruments in music—based on sustained or repeated evidence rather than isolated moments.

Reflection: Impact on Reliability and Human Alignment

Overall, aggregation improved the reliability of the system. The increase in both Micro and Macro F1 scores demonstrates that predictions became more accurate across frequent and rare instruments alike. More importantly, aggregation shifted the model's behavior from reacting to instantaneous acoustic cues toward forming decisions based on temporal evidence, which is more human-aligned.

Without aggregation, the system behaved reactively, making decisions based on momentary confidence spikes. With aggregation, the model behaved more conservatively and consistently, confirming an instrument's presence only when sufficient evidence accumulated over time.

Conclusion

This experiment confirms that temporal aggregation is essential in segmented audio classification. By averaging segment-level predictions:

- False positives were reduced
- Missed detections were mitigated
- Predictions became more stable and interpretable

Aggregation therefore plays a crucial role in transforming noisy segment-level outputs into robust, clip-level decisions, making the model's behavior more reliable and aligned with human perception.