Task 14: Data Segmentation Report

Objective

The objective of segmentation in this system is to transform a long, continuous audio track into fixed-length, overlapping temporal windows that align with the learning behavior of a CNN-based audio classifier. Segmentation ensures consistent input dimensionality, enables localized evidence learning, and allows robust reconstruction of track-level predictions from partial observations.

Segmentation is treated not merely as preprocessing, but as a core modeling decision that determines how the network perceives time.

Segmentation Parameters and Design Choices

The following segmentation parameters were selected and explicitly fixed:

Parameter	Value	Justification
Sampling Rate	16 kHz	Standard for music analysis; preserves harmonic content while reducing computation
Segment Length	3.0 seconds	Captures sufficient harmonic context for sustained instruments (piano, organ, violin)
Hop Length	1.5 seconds	50% overlap to stabilize predictions at segment boundaries
Overlap	50%	Ensures transient and intermittent instruments are not missed

These values balance temporal sensitivity and harmonic completeness, which is critical for polyphonic instrument recognition.

Implementation Strategy

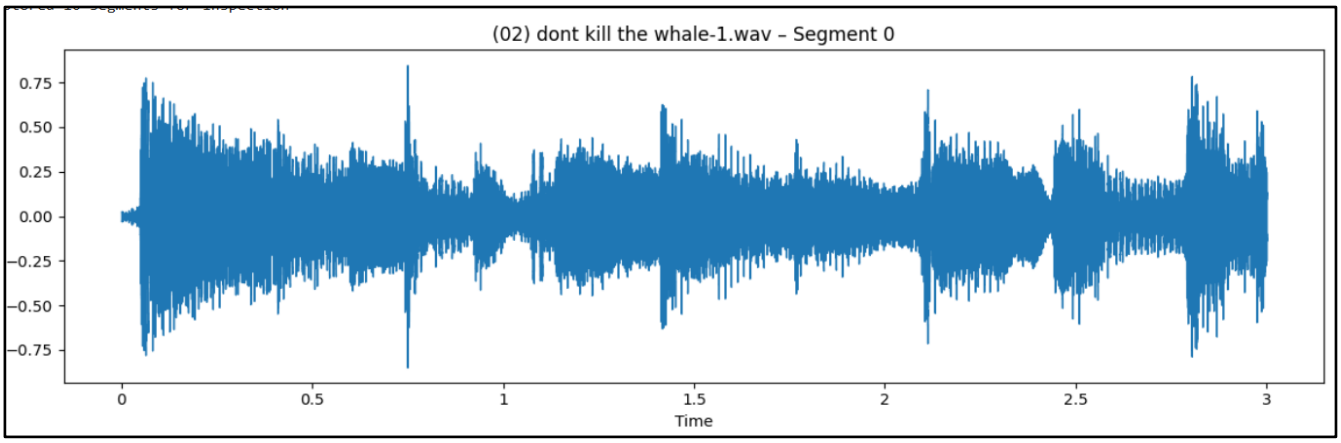
Segmentation is applied after loading the raw waveform and before spectrogram generation, following the correct theoretical pipeline:

1. Load full audio track
2. Convert stereo to mono
3. Peak normalize amplitude
4. Trim leading and trailing silence
5. Slide fixed-length windows across the waveform
6. Pad segments shorter than the target duration
7. Generate log-mel spectrograms per segment

Segments are generated in memory for inference, while a subset is saved to disk only for manual inspection, as required by Task 14.

Visual Inspection and Evidence of Correct Segmentation

4.1 Waveform Analysis (Segment 0)



The waveform of the first 3-second segment shows:

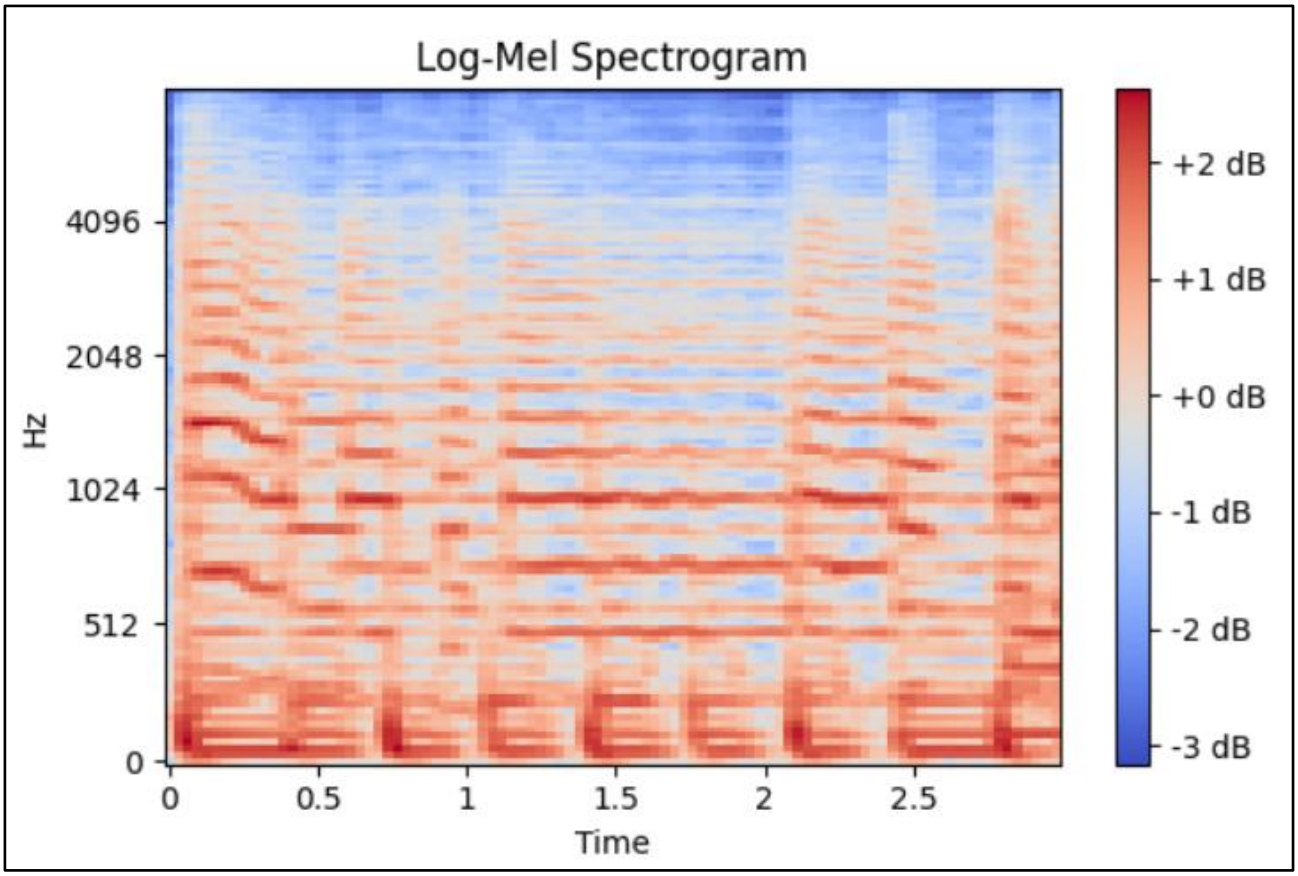
- Continuous, non-silent signal content across the full duration
- No abrupt truncation at segment boundaries
- Consistent amplitude normalization
- Clear temporal variation, indicating active musical content

This confirms that:

- Silence trimming is effective
- Segment duration is correctly enforced
- No unintended clipping or zero-padding dominates the segment

The segment is representative and information-dense, which is essential for effective CNN learning.

4.2 Log-Mel Spectrogram Analysis (Segment 0)



The corresponding log-mel spectrogram exhibits:

- Well-defined horizontal harmonic bands
- Stable energy distribution across time
- Strong low-frequency components (likely bass or piano fundamentals)
- Clear mid-frequency harmonic structure (indicative of melodic instruments)

This confirms that:

- The 3-second window preserves harmonic richness
- The mel resolution (128 bands) is sufficient
- Spectral patterns are stationary within the segment

Such time-frequency structure is exactly what convolutional kernels are designed to learn.

Alignment with CNN Learning Behaviour

CNNs do not process entire recordings holistically. Instead, they learn from localized, stationary patterns within fixed receptive fields.

The demonstrated segmentation strategy ensures:

- Fixed input size, enabling stable convolutional operations
- Local evidence extraction, allowing weak or transient instruments to be learned
- Temporal robustness, since the model learns to recognize instruments from partial context
- Boundary stability, due to overlapping windows

The spectrogram visualization clearly shows that each segment provides a self-contained, learnable representation, validating alignment with CNN inductive bias.

Role of Saved Segments

Saved audio segments serve only diagnostic and verification purposes:

- Manual listening confirms segment quality
- Visual inspection validates preprocessing correctness
- Demonstrates intentional segmentation design

Saved segments are not used for testing or evaluation, as ground-truth labels exist only at the track level. During inference, segmentation occurs in memory, and predictions are aggregated across all segments to form a track-level decision.

Final Justification

Based on the visual and auditory inspection:

- The chosen segment length preserves both harmonic and temporal information
- Overlap prevents loss of transient events
- Preprocessing ensures segments are information-dense and consistent
- Segment-level spectrograms align with CNN learning requirements

Segmentation is therefore intentional, justified, and functionally aligned with the learning behavior of the model.

Conclusion

Task 14 is successfully satisfied.
Segmentation is not treated as a passive preprocessing step, but as a deliberate design mechanism that governs temporal perception in the model. Visual and auditory inspection confirms that segmentation parameters are appropriate, correctly implemented, and supportive of robust multi-label instrument recognition.