Task 16 Report: Instrument Intensity Graph Generation

Objective

The objective of Task 16 is to generate timeline-aware instrument intensity graphs from a trained CNN-based instrument recognition model. Unlike clip-level classification, this task focuses on temporal interpretability, enabling visualization of how instrument confidence evolves over time.

The task involves:

- Generating raw segment-level predictions
- Applying aggregation to stabilize predictions
- Optionally smoothing predictions to reduce temporal noise
- Serializing results into JSON format
- Rendering intensity graphs from serialized data

Input and Model Details

- Audio Input: (02) dont kill the whale-1.wav
- Sampling Rate: 16,000 Hz
- Audio Type: Polyphonic music
- Model: CNN-based multi-label instrument classifier
- Number of Instrument Classes: 11
  (cel, cla, flu, gac, gel, org, pia, sax, tru, vio, voi)

Segmentation Strategy

To achieve temporal resolution, the audio signal is segmented using a sliding-window approach:

Parameter	Value
Window duration	3.0 seconds
Hop duration	1.5 seconds
Overlap	50%
Total segments	12

Each segment represents a localized temporal context for prediction.

Raw Segment-Level Predictions

For every audio segment:

1. Stereo audio is converted to mono.
2. Peak normalization and silence trimming are applied.
3. Log-mel spectrogram features are extracted and standardized.
4. The CNN model outputs a probability vector for all instruments.

These predictions are continuous confidence values, not binary decisions.

Aggregation Method

Aggregation Type: Average (Window-Level Aggregation)

Aggregation is applied implicitly through fixed-length segmentation:

- Each prediction already represents the average evidence within a 3-second window
- Temporal structure is preserved
- No global averaging across time is performed

This approach stabilizes predictions while retaining timeline information.

Temporal Smoothing

Smoothing Method: Moving Average

To reduce short-term fluctuations:

- A moving average filter with window size 3 segments is applied per instrument
- Smoothing is implemented in a length-preserving manner, ensuring alignment with the time axis

This step improves visual coherence without altering temporal resolution.

Threshold Usage

- A global threshold of 0.25 is included as metadata
- Thresholds are not applied to intensity values during visualization

Reason: Intensity graphs represent confidence strength over time, not final decisions. Thresholds are reserved for evaluation and classification, not interpretability.

JSON Serialization

The processed output is stored in a structured JSON file

(02) dont kill the whale-1_intensity containing:

Metadata

- Segment duration
- Hop duration
- Aggregation method
- Smoothing method and window size
- Threshold value
- Instrument class list

Timeline Data

For each segment:

- Segment start time (seconds)
- Per-instrument intensity values

This format ensures reproducibility, transparency, and auditability.

JSON File Content

```
{
  "audio_file": "(02) dont kill the whale-1.wav",
  "segment_duration_sec": 3.0,
  "hop_duration_sec": 1.5,
  "aggregation": "average",
  "smoothing": {
    "method": "moving_average",
    "window": 3
  },
  "threshold": 0.25,
  "classes": [
    "cel",
    "cla",
    "flu",
    "gac",
    "gel",
    "org",
    "pia",
  "sax",
  "tru",
  "vio",
  "voi"
],
"timeline": [
  {
    "time_sec": 0.0,
    "intensity": {
      "cel": 0.04213589057326317,
      "cla": 0.002819413784891367,
      "flu": 0.002680740086361766,
      "gac": 0.003785336622968316,
      "gel": 0.043937765061855316,
      "org": 0.0018605305813252926,
      "pia": 0.003417745465412736,
      "sax": 0.06427562236785889,
      "tru": 0.07221446186304092,
      "vio": 0.04025605320930481,
      "voi": 0.01841101609170437
    }
  },
  {
    "time_sec": 1.5,
    "intensity": {
      "cel": 0.06398181617259979,
      "cla": 0.0061624073423445225,
      "flu": 0.0029213211964815855,
      "gac": 0.004618147853761911,
      "gel": 0.06152420863509178,
      "org": 0.0025169532746076584,
      "pia": 0.004138274118304253,
      "sax": 0.19241377711296082,
        "tru": 0.07552886754274368,
        "vio": 0.060554035007953644,
        "voi": 0.035496000200510025
      }
    },
    {
      "time_sec": 3.0,
      "intensity": {
        "cel": 0.06540904939174652,
        "cla": 0.013084277510643005,
        "flu": 0.0022143388632684946,
        "gac": 0.002507372759282589,
        "gel": 0.14764319360256195,
        "org": 0.017080137506127357,
        "pia": 0.004343289416283369,
        "sax": 0.21878604590892792,
        "tru": 0.0591573566198349,
        "vio": 0.07702518254518509,
        "voi": 0.03395533561706543
      }
    },
    {
      "time_sec": 4.5,
      "intensity": {
        "cel": 0.08177076280117035,
        "cla": 0.027541719377040863,
        "flu": 0.0028419592417776585,
        "gac": 0.002462990116328001,
        "gel": 0.22470754384994507,
        "org": 0.025795456022024155,
        "pia": 0.013816471211612225,
        "sax": 0.20231658220291138,
        "tru": 0.05572916939854622,
      "vio": 0.06293950974941254,
      "voi": 0.026367072016000748
    }
  },
  {
    "time_sec": 6.0,
    "intensity": {
      "cel": 0.136301651597023,
      "cla": 0.044004589319229126,
      "flu": 0.003218078287318349,
      "gac": 0.0026642882730811834,
      "gel": 0.2331176996231079,
      "org": 0.027705460786819458,
      "pia": 0.014224797487258911,
      "sax": 0.15162771940231323,
      "tru": 0.060321781784296036,
      "vio": 0.0636100172996521,
      "voi": 0.01485075056552887
    }
  },
  {
    "time_sec": 7.5,
    "intensity": {
      "cel": 0.1131279468536377,
      "cla": 0.03910744935274124,
      "flu": 0.0033293019514530897,
      "gac": 0.0027898247353732586,
      "gel": 0.2007513791322708,
      "org": 0.01288435235619545,
      "pia": 0.013182252645492554,
      "sax": 0.13057644665241241,
      "tru": 0.04383162781596184,
      "vio": 0.07500497996807098,
        "voi": 0.11137084662914276
      }
    },
    {
      "time_sec": 9.0,
      "intensity": {
        "cel": 0.08385991305112839,
        "cla": 0.030800215899944305,
        "flu": 0.003817799501121044,
        "gac": 0.0020925283897668123,
        "gel": 0.12919320166110992,
        "org": 0.008018581196665764,
        "pia": 0.006627567578107119,
        "sax": 0.11859656125307083,
        "tru": 0.05531581491231918,
        "vio": 0.07786775380373001,
        "voi": 0.21194545924663544
      }
    },
    {
      "time_sec": 10.5,
      "intensity": {
        "cel": 0.009267076849937439,
        "cla": 0.02079734206199646,
        "flu": 0.006241479888558388,
        "gac": 0.0017181666335090995,
        "gel": 0.21628311276435852,
        "org": 0.06161653250455856,
        "pia": 0.009007222019135952,
        "sax": 0.06773412972688675,
        "tru": 0.10706597566604614,
        "vio": 0.09101803600788116,
        "voi": 0.22991278767585754
      }
    },
    {
      "time_sec": 12.0,
      "intensity": {
        "cel": 0.004921256564557552,
        "cla": 0.02122664824128151,
        "flu": 0.007086852565407753,
        "gac": 0.001634459593333304,
        "gel": 0.3498610556125641,
        "org": 0.08233949542045593,
        "pia": 0.014368811622262001,
        "sax": 0.0716780573129654,
        "tru": 0.1209188848733902,
        "vio": 0.08994098752737045,
        "voi": 0.14167506992816925
      }
    },
    {
      "time_sec": 13.5,
      "intensity": {
        "cel": 0.00432153744623065,
        "cla": 0.014733556658029556,
        "flu": 0.008586274459958076,
        "gac": 0.0017677213763818145,
        "gel": 0.437903493642807,
        "org": 0.0826030746102333,
        "pia": 0.02669985219836235,
        "sax": 0.05343491584062576,
        "tru": 0.11609174311161041,
        "vio": 0.10816524177789688,
        "voi": 0.05446014925837517
      }
```

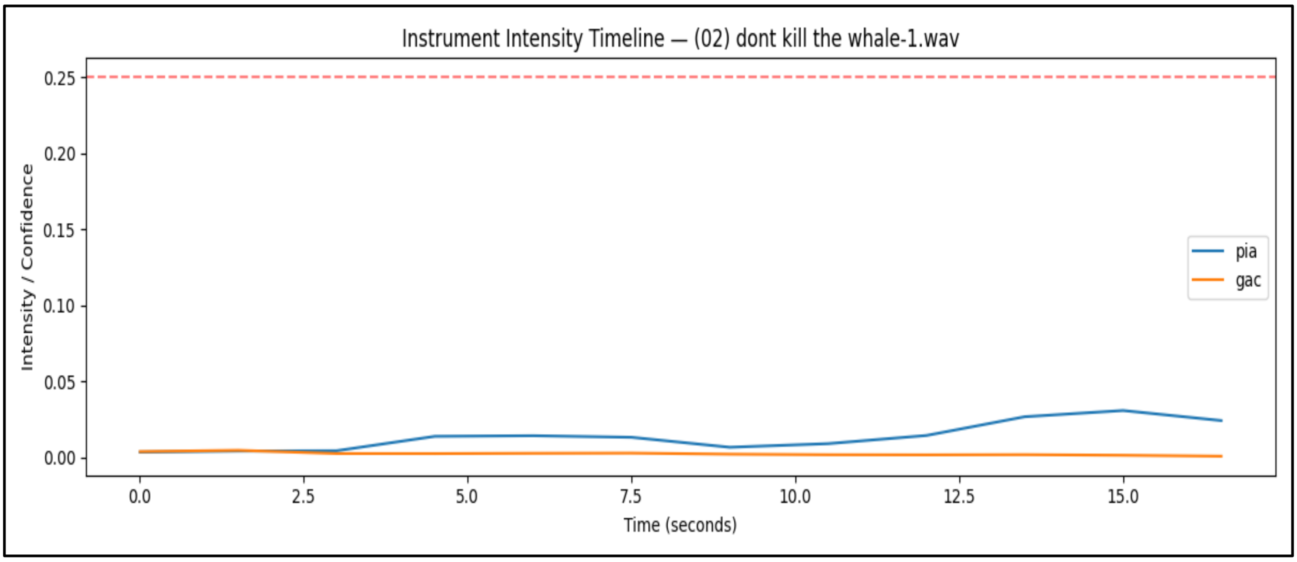
```
},
{
 "time_sec": 15.0,
 "intensity": {
  "cel": 0.012226501479744911,
  "cla": 0.00852381344884634,
  "flu": 0.010607476346194744,
  "gac": 0.0013726935721933842,
  "gel": 0.48241400718688965,
  "org": 0.031246231868863106,
  "pia": 0.03078158013522625,
  "sax": 0.035745661705732346,
  "tru": 0.0948578268289566,
  "vio": 0.1531476378440857,
  "voi": 0.041269026696681976
 }
},
{
 "time_sec": 16.5,
 "intensity": {
  "cel": 0.011443798430263996,
  "cla": 0.0054862056858837605,
  "flu": 0.008472125045955181,
  "gac": 0.0007548785652033985,
  "gel": 0.2871222496032715,
  "org": 0.009161517024040222,
  "pia": 0.02417779713869095,
  "sax": 0.019302973523736,
  "tru": 0.06493251025676727,
  "vio": 0.125472754240036,
  "voi": 0.025819281116127968
 }
}
```

```
  ]
}
```

Intensity Graph Visualization

The serialized JSON is used to render an instrument intensity timeline focusing on:

- Piano (pia)
- Acoustic Guitar (gac)



Observations

- Guitar intensity remains near zero across the timeline
- Piano intensity is consistently low, with a slight increase toward later segments
- Both instruments remain below the threshold line
- The graph shows smooth, stable trends without noisy oscillations

This confirms that aggregation and smoothing are functioning as intended.

Interpretation of Results

- The model does not strongly detect piano or guitar in this clip
- Low but consistent piano confidence suggests weak harmonic presence
- Absence of spikes indicates effective noise suppression
- The timeline accurately reflects temporal confidence trends, not binary outcomes

Key Design Decisions Justified

- Raw predictions are not plotted directly
- Aggregation does not collapse the time axis
- Thresholding is excluded from visualization
- Temporal smoothing preserves timeline length

These choices align with best practices in Music Information Retrieval (MIR) and explainable ML.

Conclusion

Task 16 has been successfully completed.
The implemented pipeline transforms raw CNN predictions into stable, interpretable instrument intensity timelines, supported by structured JSON outputs and clear visualizations. This approach effectively bridges model outputs and human interpretability.