

Seeing Sound: Investigating the Effects of Visualizations and Complexity on Crowdsourced Audio Annotations

Mark Cartwright¹, Ayanna Seals¹, Justin Salamon¹, Alex Williams², Stefanie Mikloska², Duncan MacConnell¹, Edith Law², Juan Bello¹, Oded Nov¹

1. New York University
2. University of Waterloo

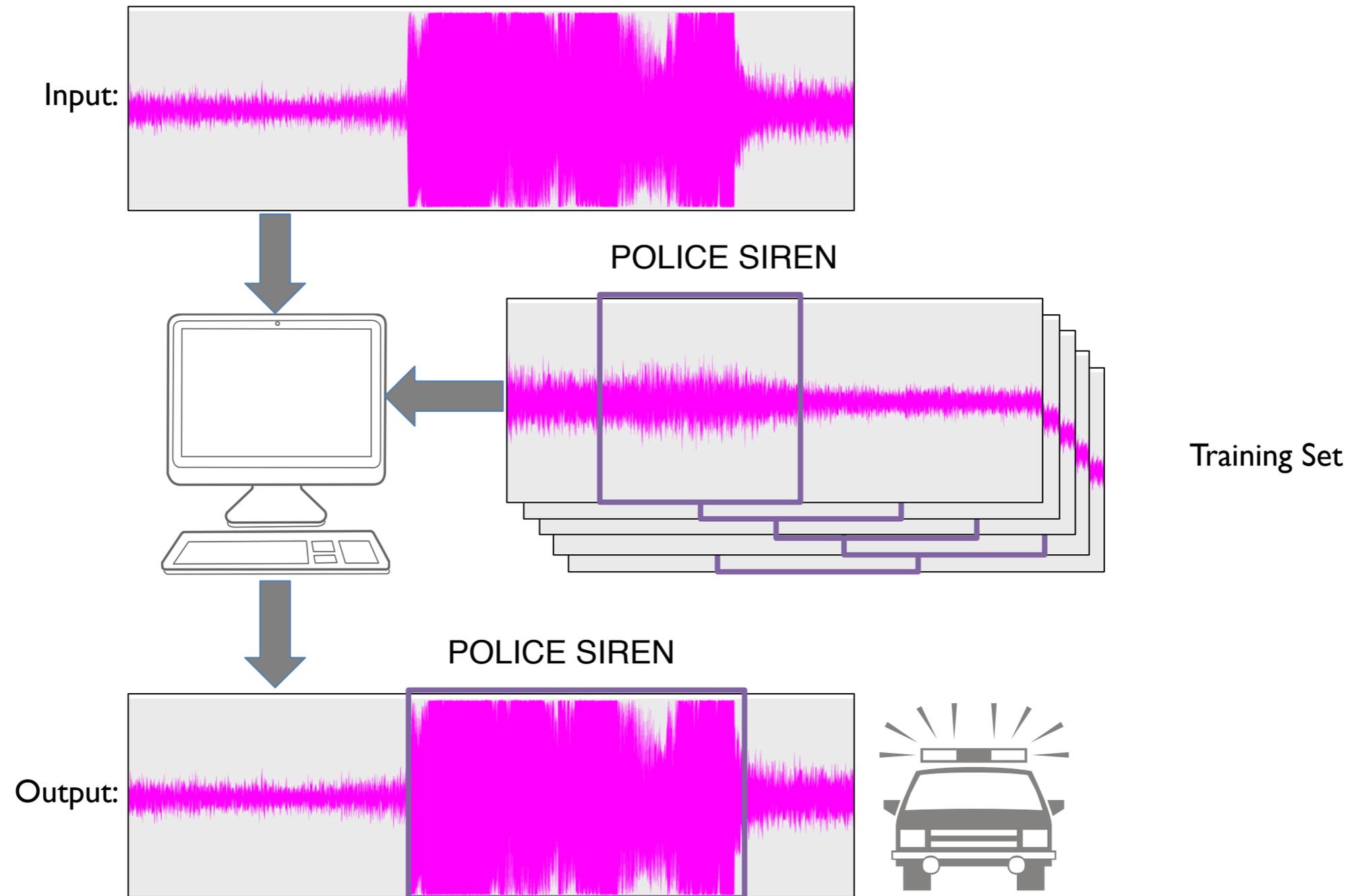




Sounds of New York City

A cyber-physical system powered by an acoustic sensor network that aims to **monitor, analyze, and mitigate** urban noise pollution.

Audio Annotation of Sound-Event Detection



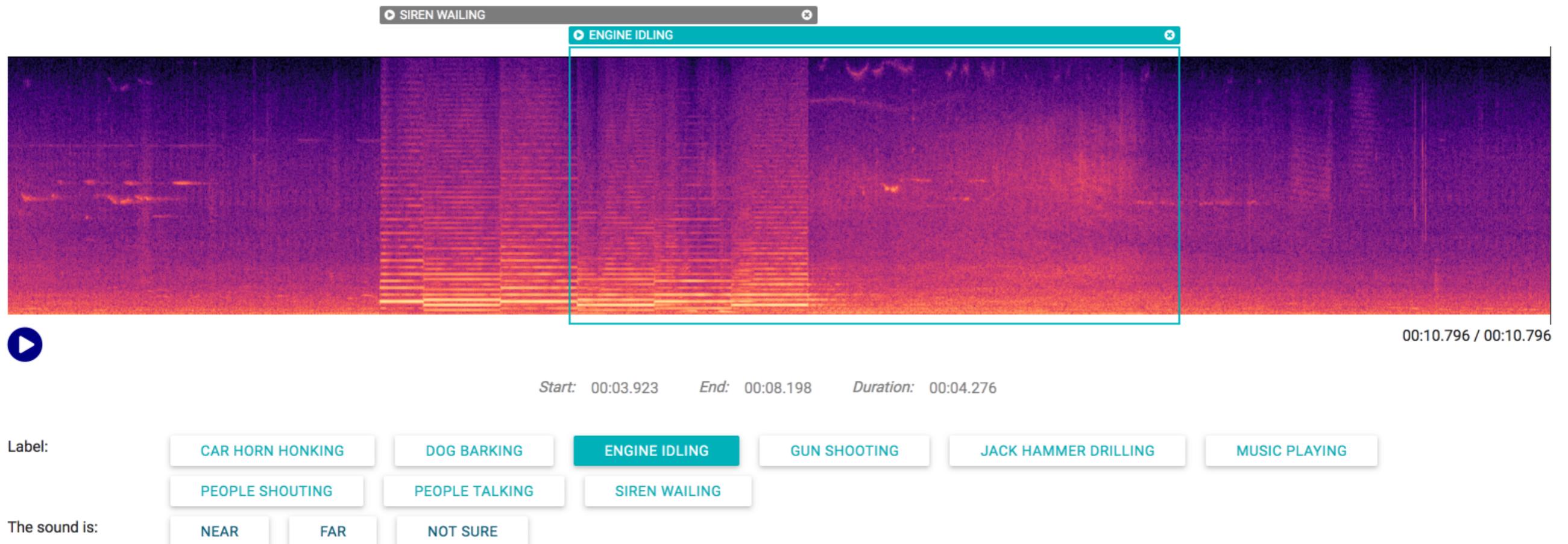
Research Questions

- Which sound visualization aid yields the highest quality crowdsourced audio annotations?
- What limitations can we expect from crowdsourced audio annotations as a function of soundscape complexity?
- What is the trade-off between reliability and redundancy in crowdsourced audio annotation?



The Audio Annotator

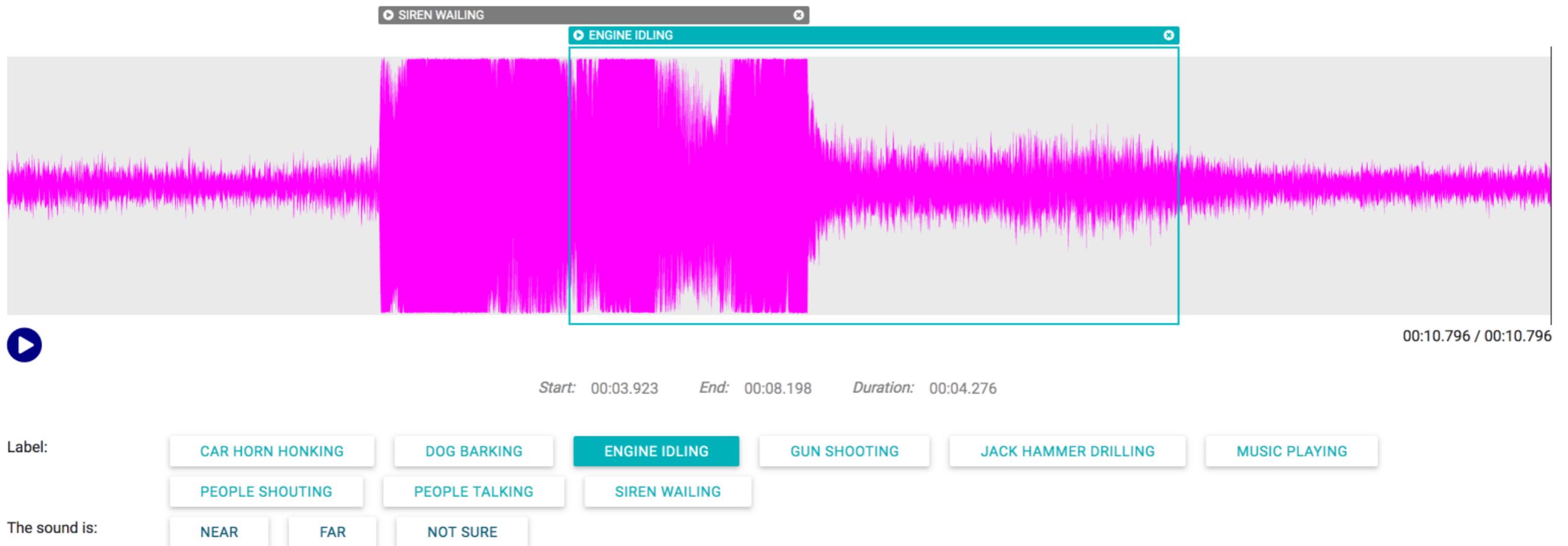
Configured with the spectrogram visualization:



github.com/CrowdCurio/audio-annotator

The Audio Annotator

Configured with the waveform visualization:



The screenshot displays the Audio Annotator interface. At the top, a waveform visualization shows a pink audio signal. Two overlapping annotations are present: a grey bar labeled 'SIREN WAILING' and a teal bar labeled 'ENGINE IDLING'. Below the waveform, a play button is on the left, and the current time '00:10.796 / 00:10.796' is on the right. The selected annotation's metadata is shown: 'Start: 00:03.923 End: 00:08.198 Duration: 00:04.276'. Underneath, there are two rows of label buttons. The first row includes 'CAR HORN HONKING', 'DOG BARKING', 'ENGINE IDLING' (highlighted in teal), 'GUN SHOOTING', 'JACK HAMMER DRILLING', and 'MUSIC PLAYING'. The second row includes 'PEOPLE SHOUTING', 'PEOPLE TALKING', and 'SIREN WAILING'. At the bottom, the 'The sound is:' section contains three buttons: 'NEAR', 'FAR', and 'NOT SURE'.

github.com/CrowdCurio/audio-annotator

The Audio Annotator

Configured without a visualization:

The screenshot displays the Audio Annotator interface. At the top, there are two floating labels: "SIREN WAILING" and "ENGINE IDLING". The "ENGINE IDLING" label is highlighted with a teal border. Below the labels is a large grey rectangular area representing the audio waveform, which is currently empty. A play button is located at the bottom left. The timeline at the bottom shows the current position at 00:10.796 / 00:10.796. The selected segment has a start time of 00:03.923, an end time of 00:08.198, and a duration of 00:04.276. The "Label:" section contains several buttons: "CAR HORN HONKING", "DOG BARKING", "ENGINE IDLING" (highlighted), "GUN SHOOTING", "JACK HAMMER DRILLING", and "MUSIC PLAYING". The "The sound is:" section contains buttons for "NEAR", "FAR", and "NOT SURE".

github.com/CrowdCurio/audio-annotator



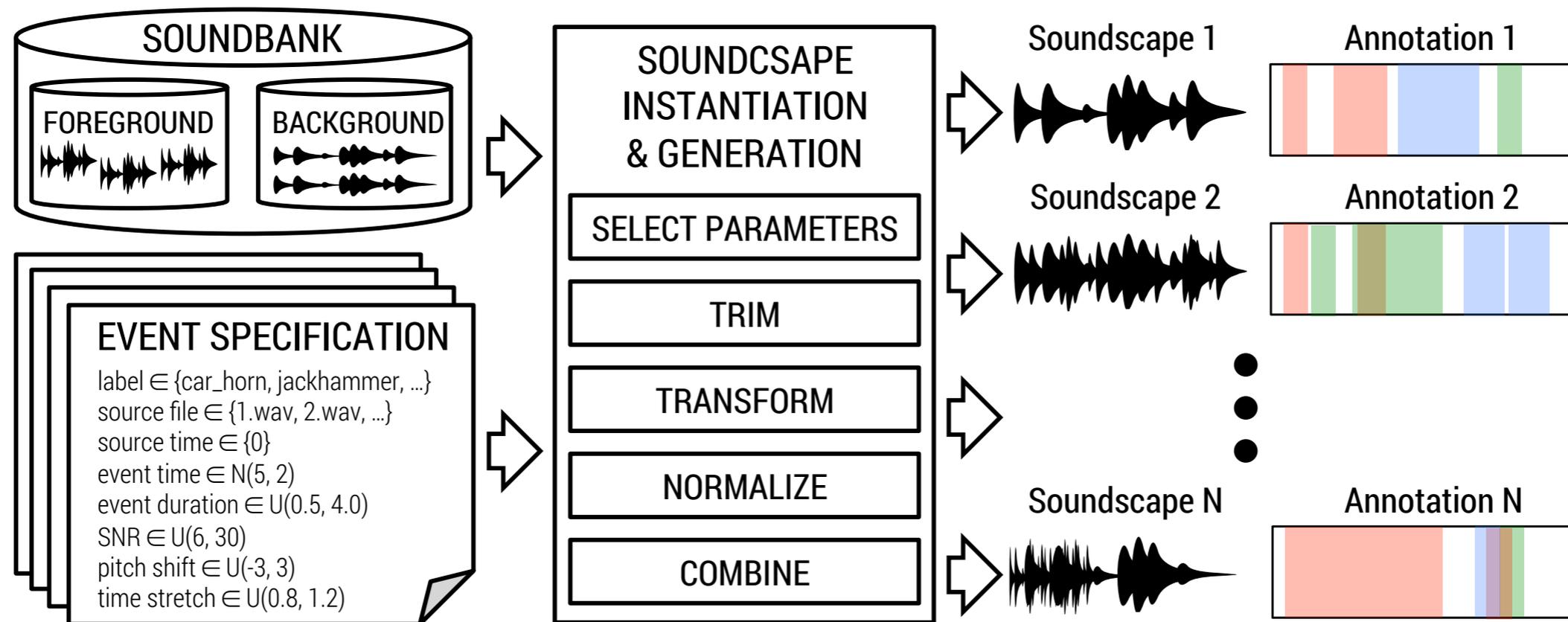
CrowdCurio.

Fostering Curiosity Through Science.

crowdcurio.com

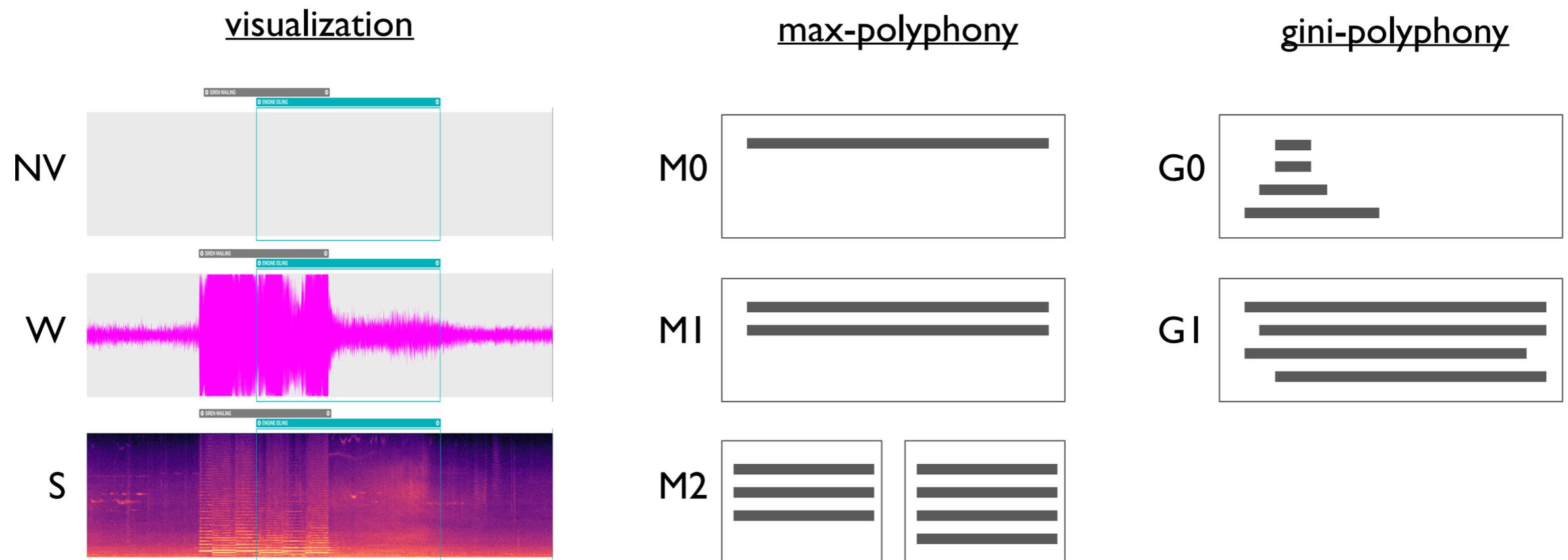
Scaper: Soundscape Synthesis

- Open source python library for soundscape synthesis (WASPAA 2017)
- github.com/justinsalamon/scaper



Experiment

- 3 x 3 x 2 between-subjects factorial design:



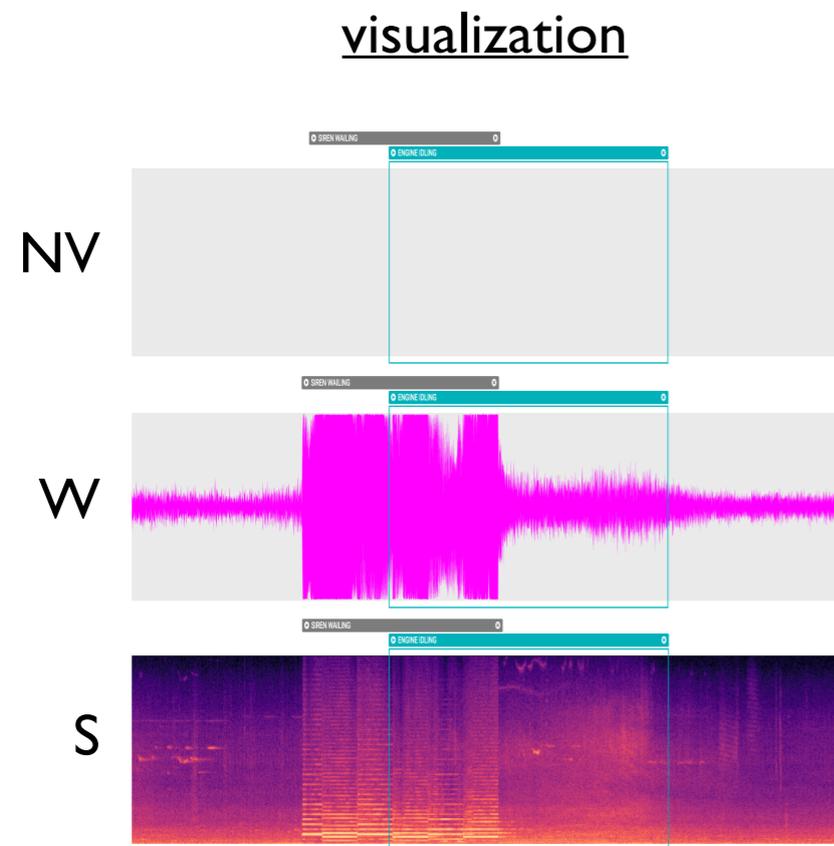
- Soundscape examples:
M0G0 M0G1

M2G0

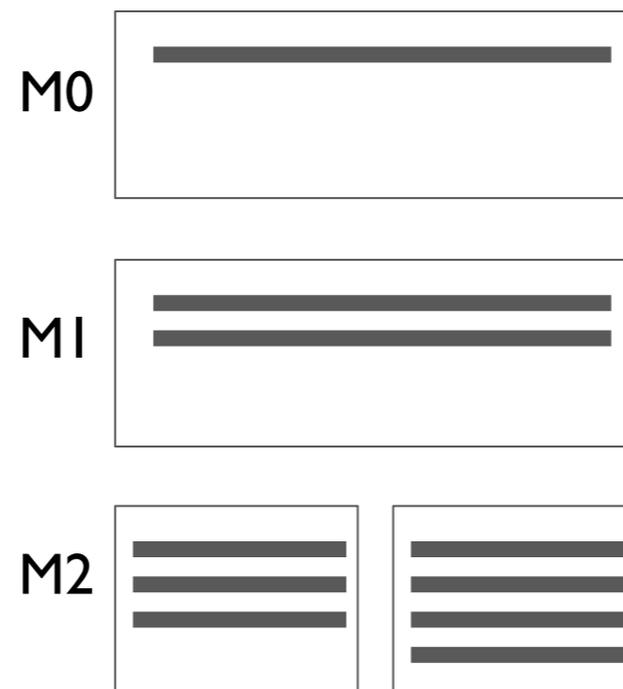
M2G1

Experiment

- 3 x 3 x 2 between-subjects factorial design:



max-polyphony



gini-polyphony



- Soundscape examples:

M0G0

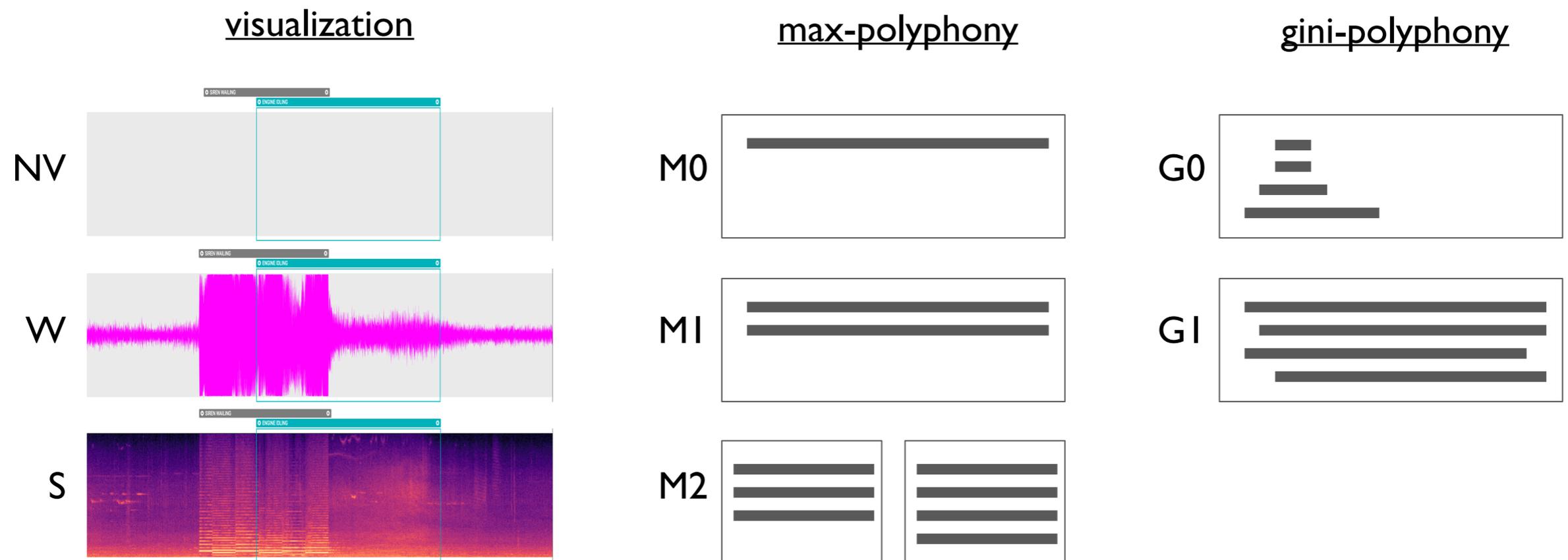
M0G1

M2G0

M2G1

Experiment

- 3 x 3 x 2 between-subjects factorial design:



- Soundscape examples:

M0G0

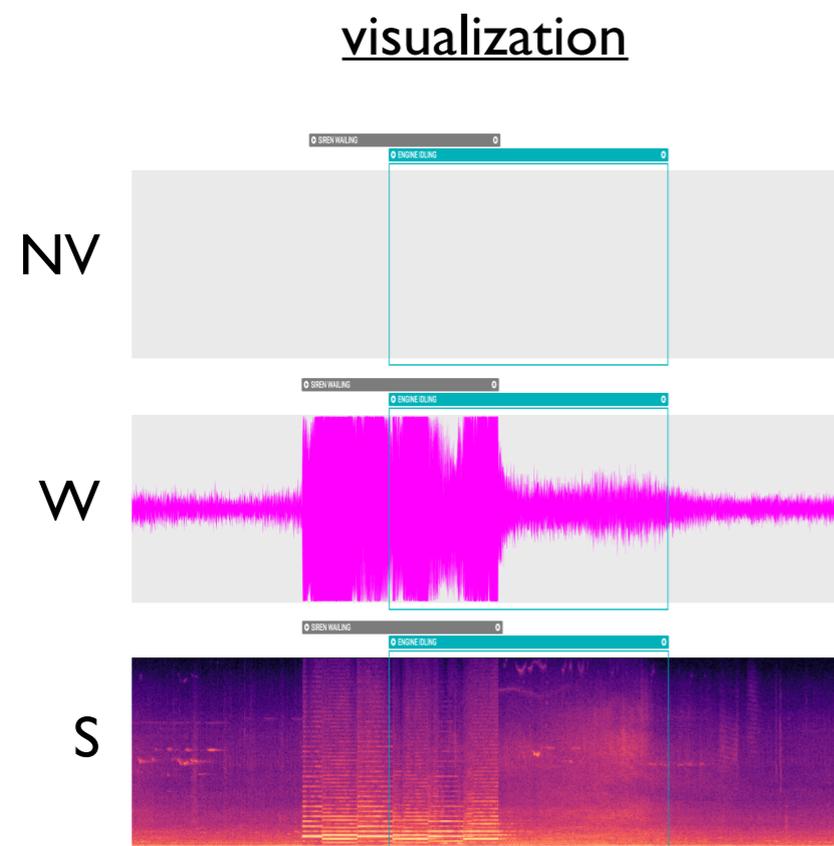
M0G1

M2G0

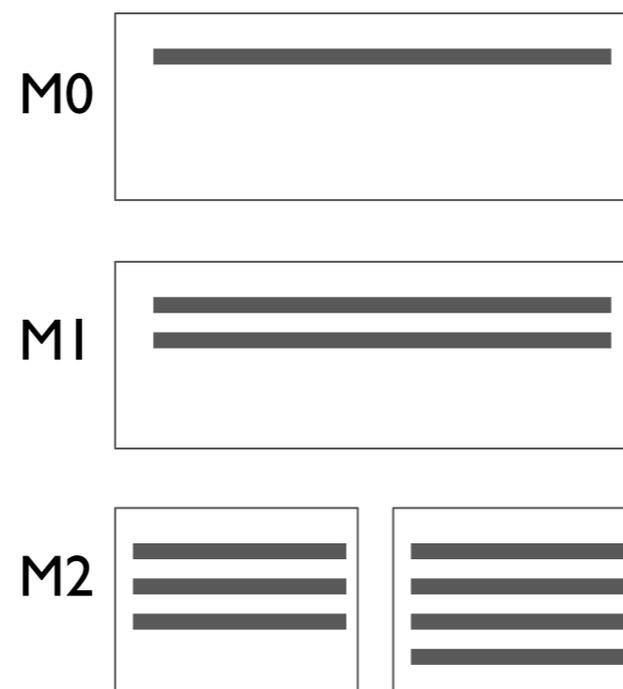
M2G1

Experiment

- 3 x 3 x 2 between-subjects factorial design:



max-polyphony



gini-polyphony



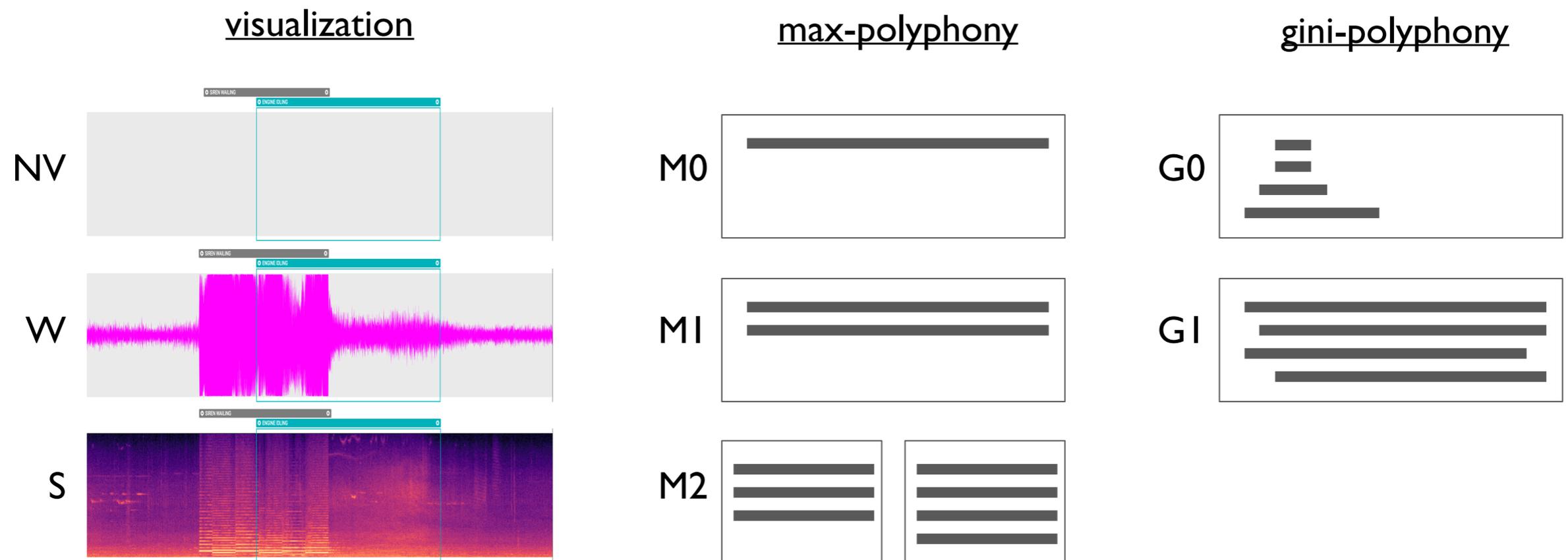
- Soundscape examples:
M0G0 M0G1

M2G0

M2G1

Experiment

- 3 x 3 x 2 between-subjects factorial design:



- Soundscape examples:
M0G0 M0G1

M2G0

M2G1

Experiment

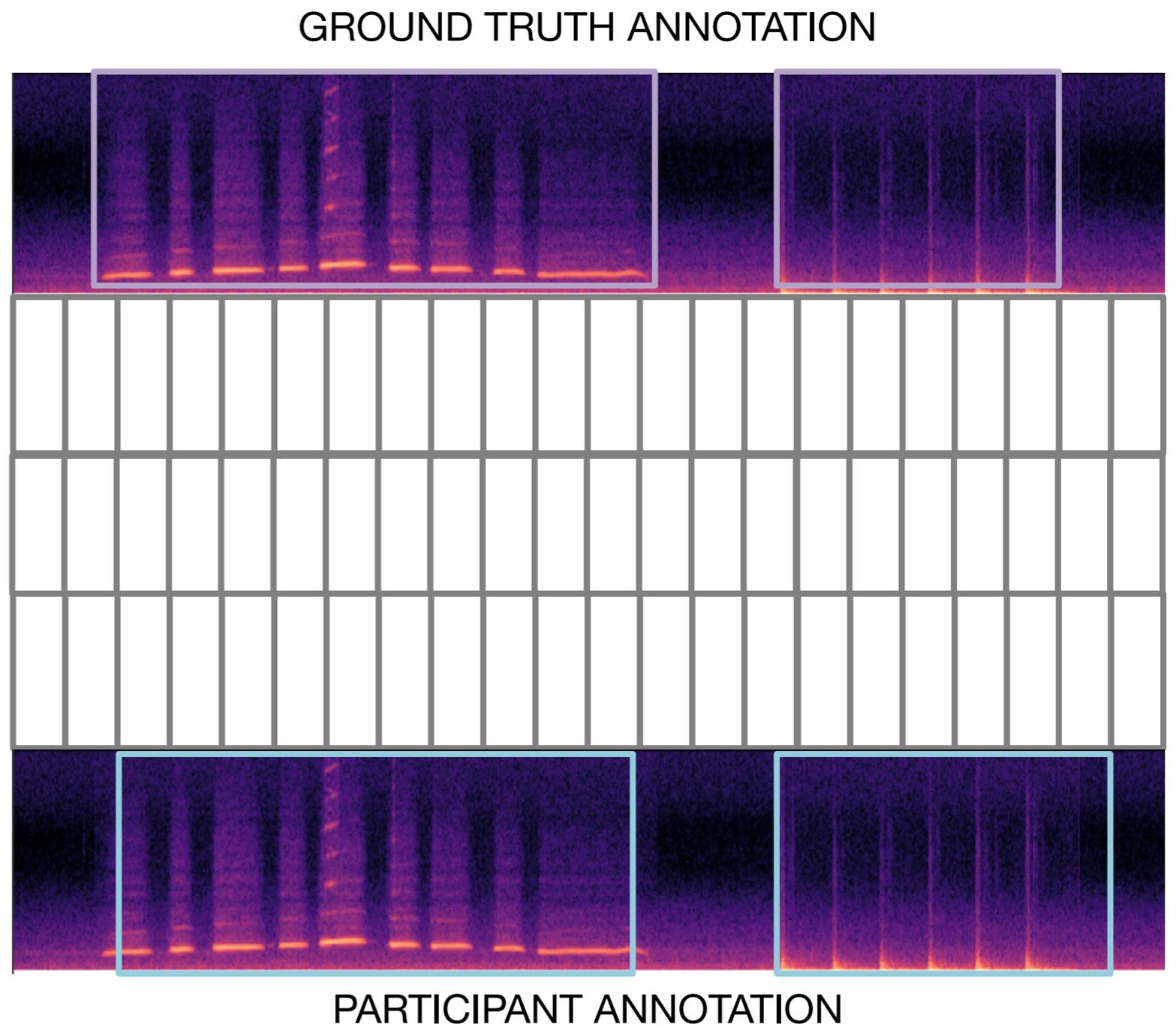
- 10 s synthesized urban soundscapes (i.e. audio stimuli)
- Classes: *car horn honking, dog barking, engine idling, gun shooting, jack hammer drilling, music playing, people shouting, people talking, siren wailing*
- 30 replications / 540 participants from Mechanical Turk
- 10 soundscapes per complexity condition (i.e. max- x gini-polyphony pair)
- Counterbalanced ordering of soundscapes
- Ran on the CrowdCurio platform

Participant Tasks

- Hearing screening
- Pre-task questionnaire
- Tutorial video
- Practice annotation task
- Series of 10 annotation tasks
- Post-task questionnaire

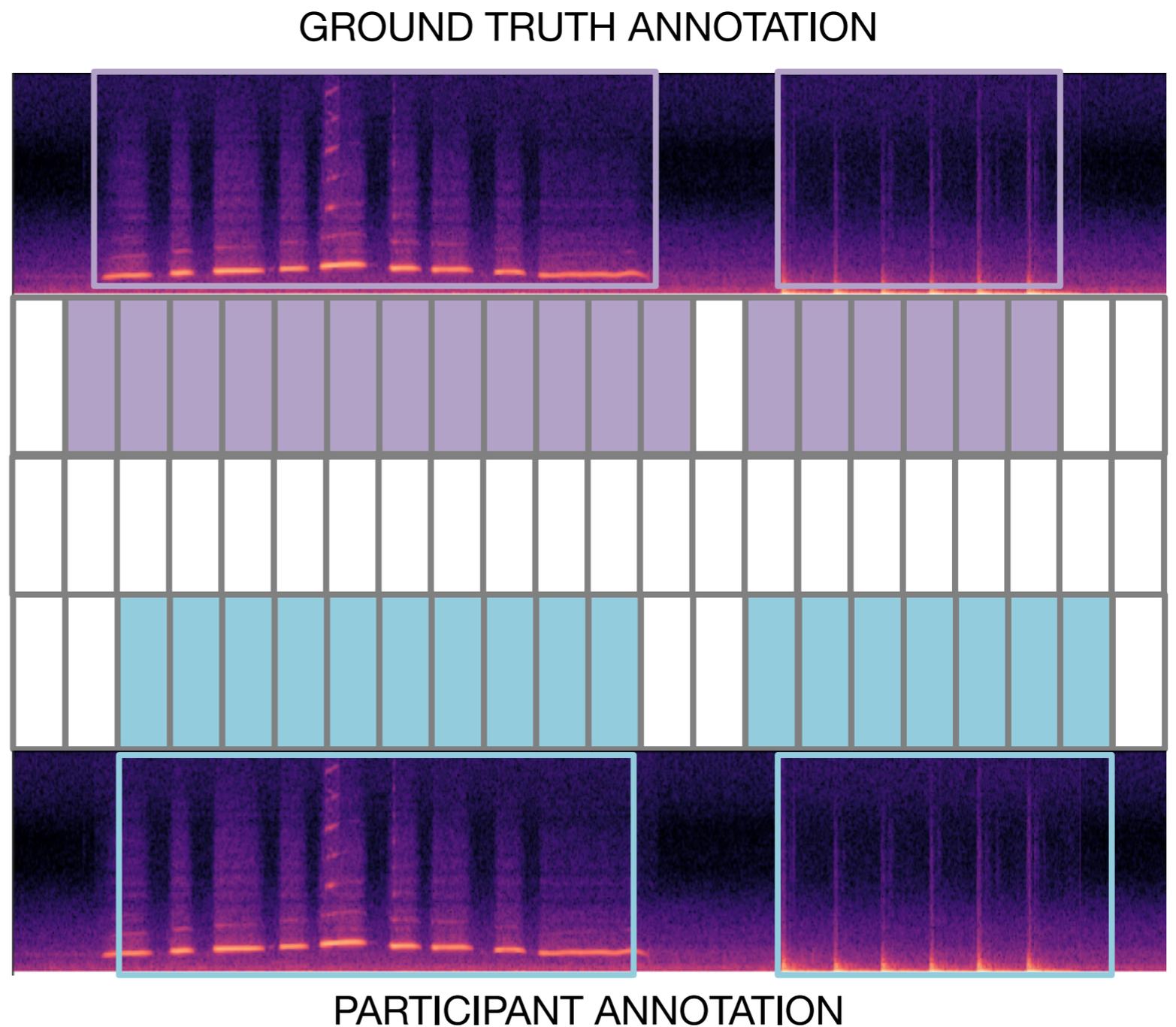
Frame-based Evaluation

- Segment signal into 100ms frames.



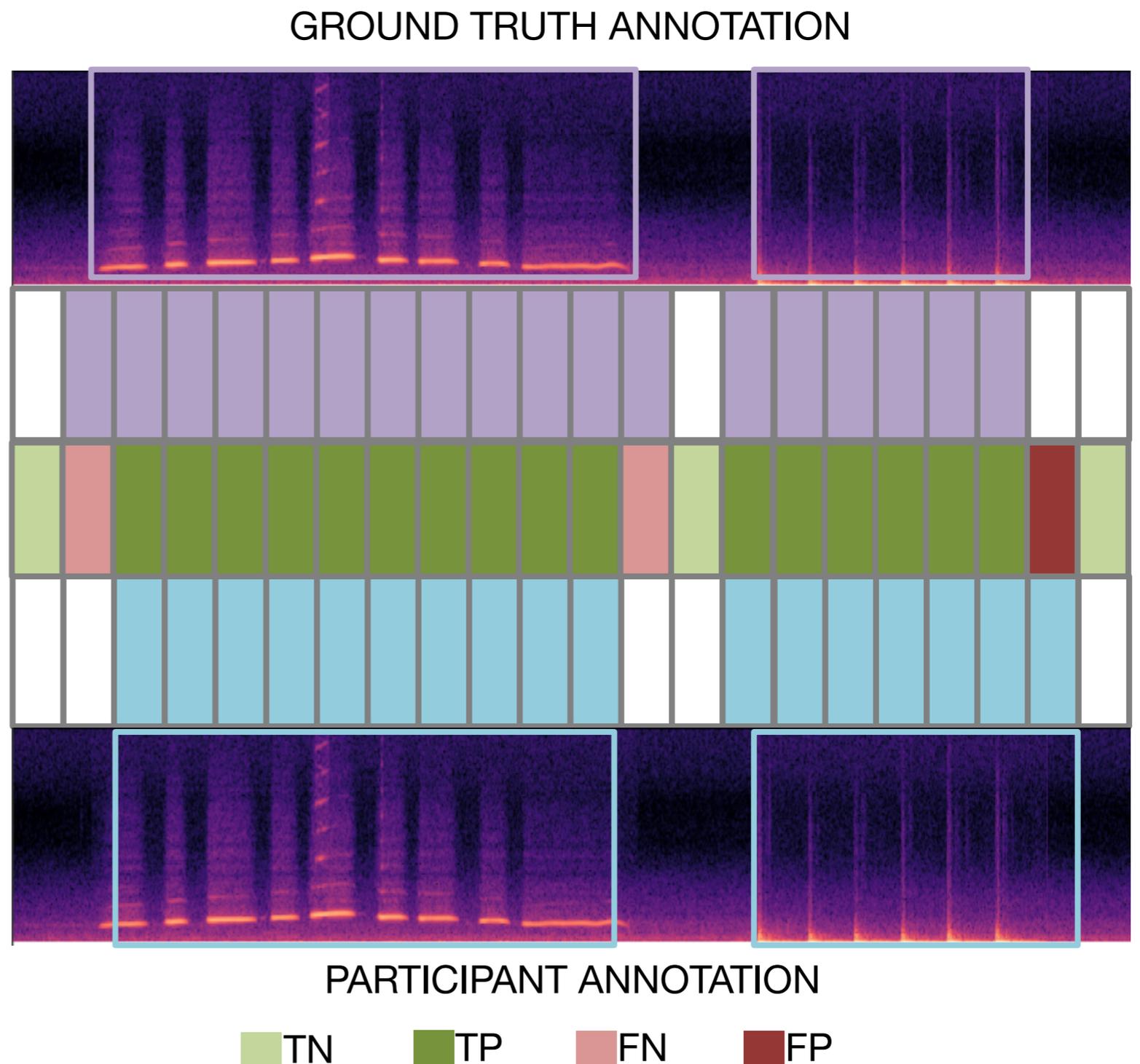
Frame-based Evaluation

- Segment signal into 100ms frames.
- Round the annotations to the outer frame boundaries



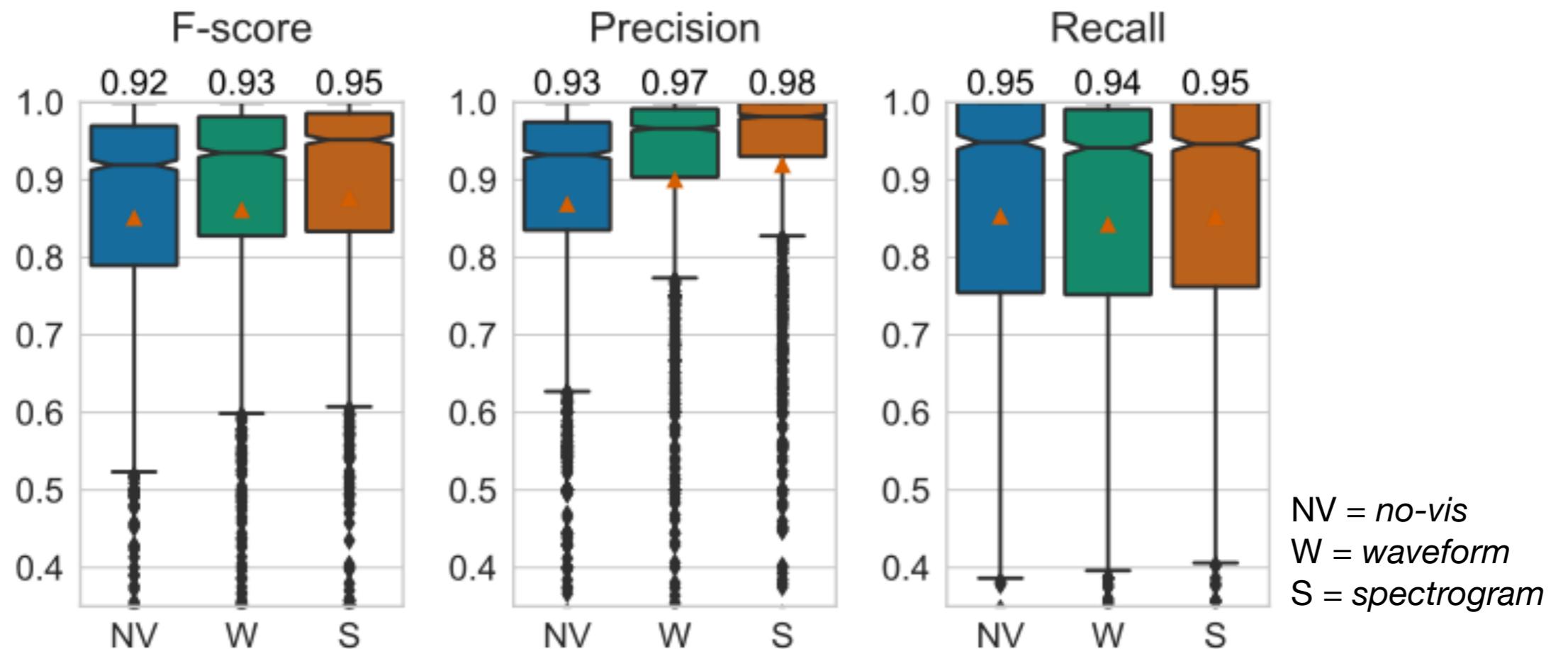
Frame-based Evaluation

- Segment signal into 100ms frames.
- Round the annotations to the outer frame boundaries
- Count TP, FP, FN for each class and calculate precision, recall, F-score



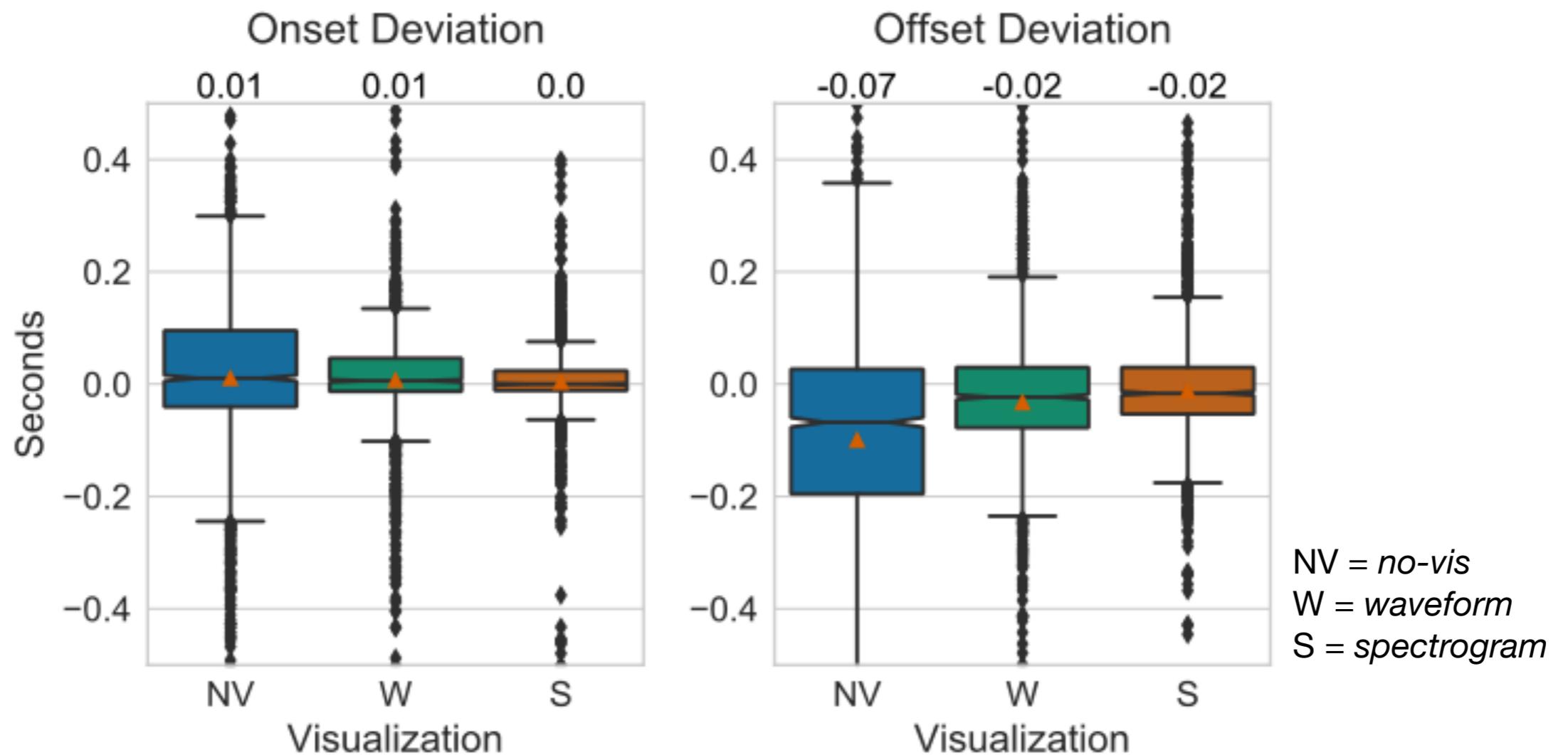
Results

Effect of Visualization on Quality of Annotations

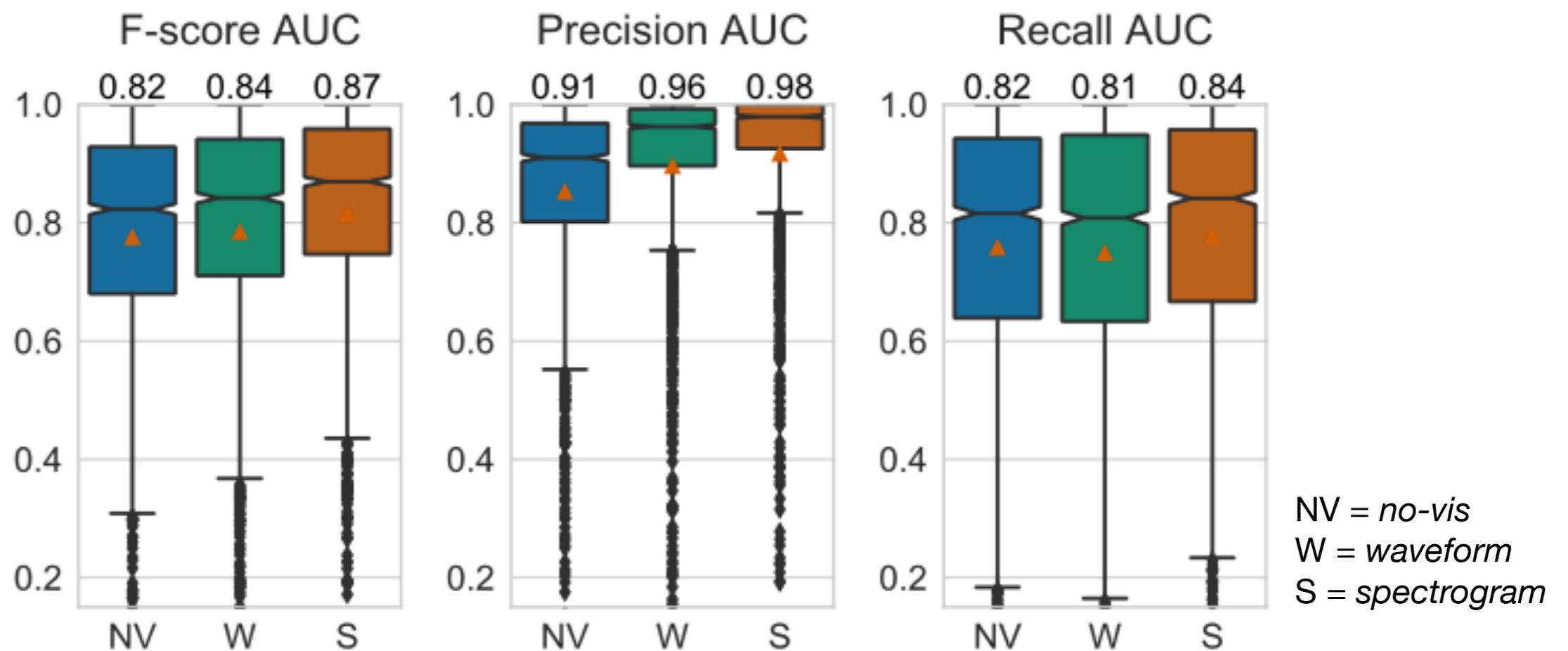


Spectrogram → higher-quality annotations

Effect of Visualization on Quality of Annotations

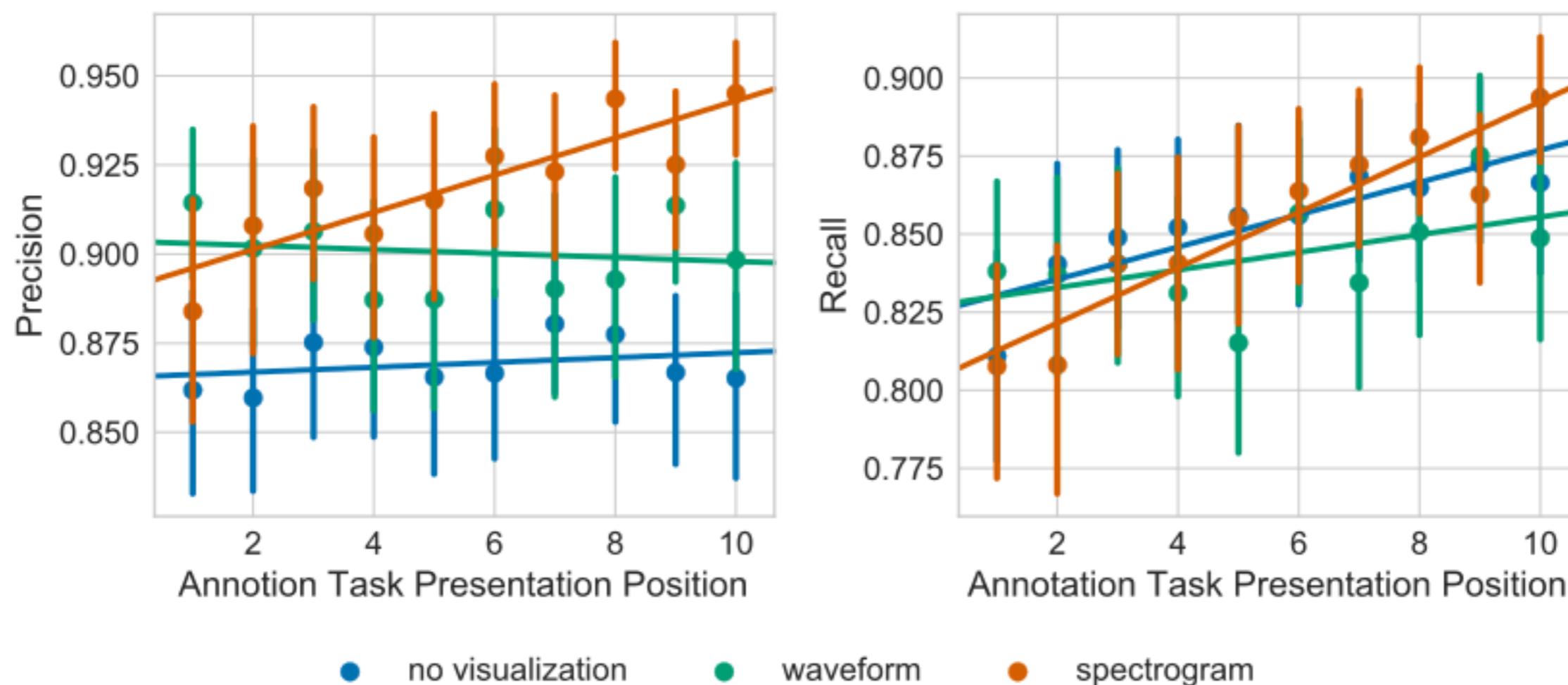


Effect of Visualization on Quality and Speed of Annotations



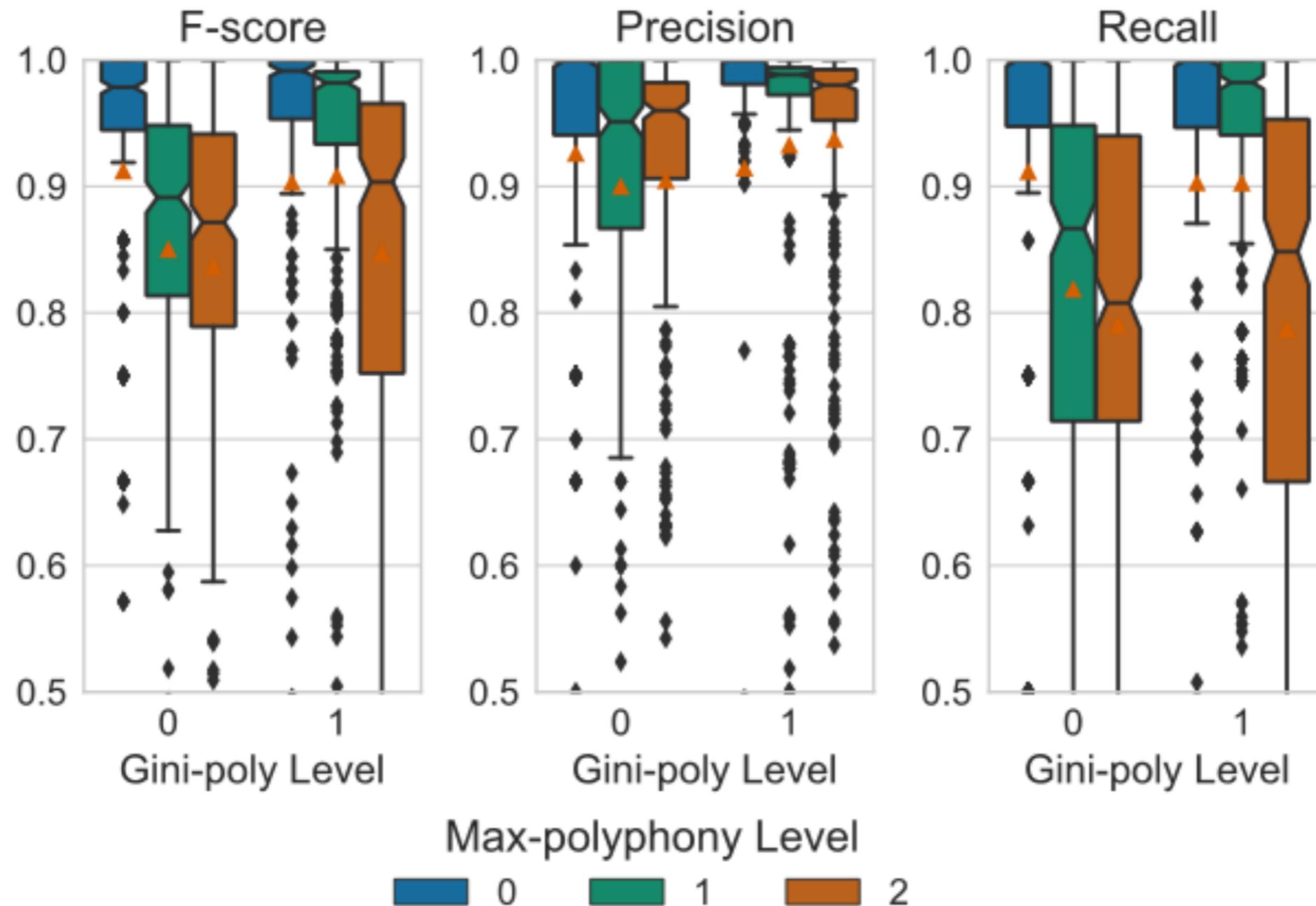
Spectrogram → higher-quality and faster annotations

Effect of Visualization on Task Learning



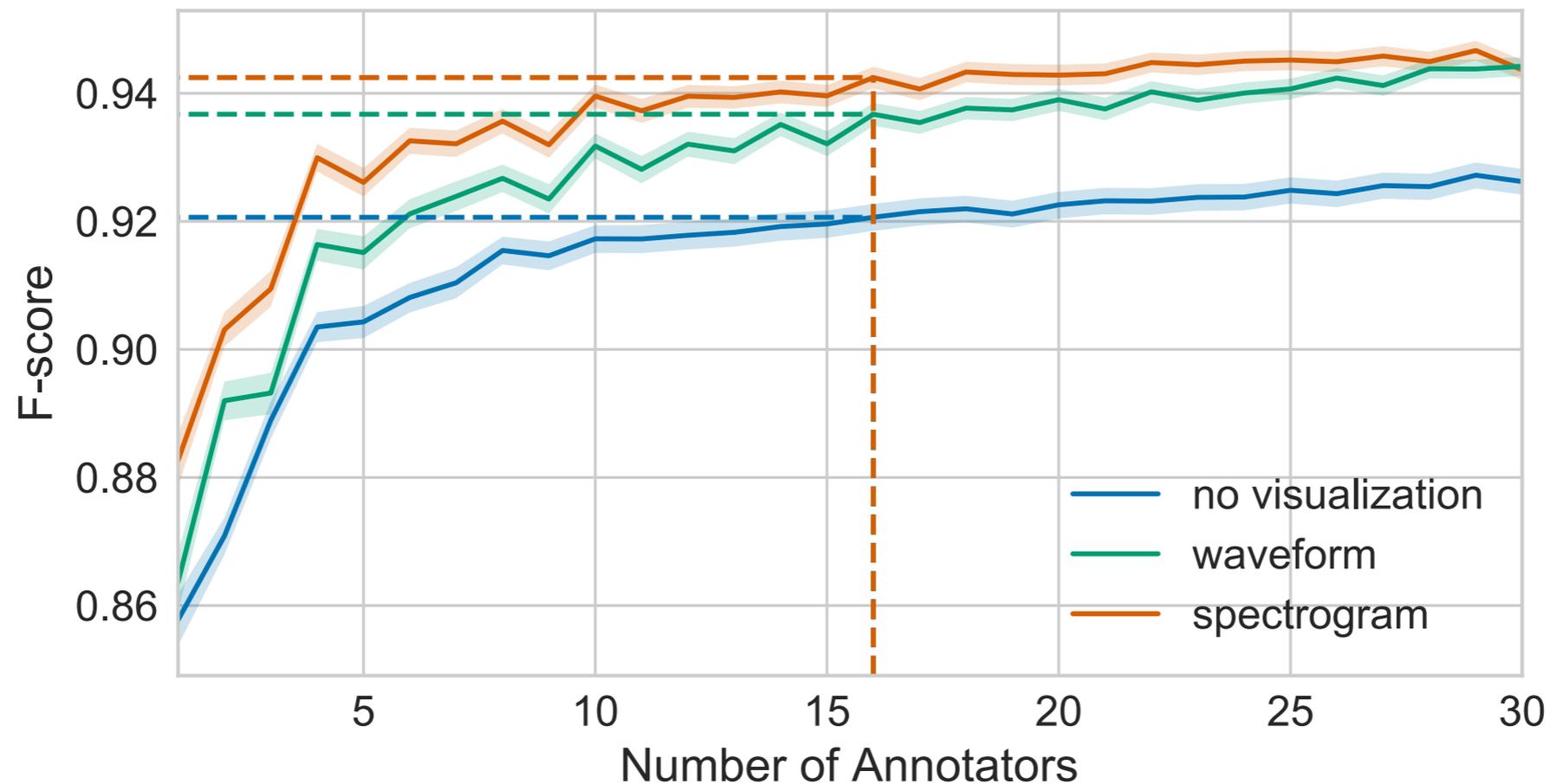
Expect even higher quality annotations after learning period

Effect of Soundscape Complexity on Annotation Quality



Complex soundscapes → expect precise but incomplete annotations

Effect of Number of Annotators on Aggregate Annotation Quality



16 annotators captured 90% of gain in annotation quality, but 5 annotators is reasonable choice with respect to cost/quality trade-off

Takeaways

- Spectrogram → higher-quality and faster annotations
 - Expect even higher quality annotations after learning period
 - Complex soundscapes → expect precise but incomplete annotations
 - 5 annotators is reasonable choice with respect to cost/quality trade-off
-

SONYC: wp.nyu.edu/sonyc

Audio Annotator: github.com/CrowdCurio/audio-annotator

Scaper: github.com/justinsalamon/scaper

CrowdCurio: crowdcurio.com

Data: <https://doi.org/10.5281/zenodo.887924>