Standards to describe audio event sequences/segments

Dan Stowell
Machine Listening Lab, Centre for Digital Music, QMUL, London
(c) 2020-06-08
for TDWG Audubon Core Maintenance Group

Requirement

In bioacoustics there is commonly a need to annotate/segment audio recordings in terms of the temporal phenomena heard within. The annotation takes various forms, e.g.:
(a) a list of instantaneous point events;
(b) nonoverlapping time-regions or "segments";
(c) potentially-overlapping regions each with its own beginning and end time;
(d) same as c but also with lower and upper frequency bounds per region.
A given annotation could focus on one taxon or individual, or could attempt complete coverage of all sound events heard. Regions (or even sets of regions ["tracks"]) could be labelled by taxon, but could also be labelled with non-taxon terms (e.g. *unknown, background noise, car*) or within-taxon designations (e.g. *call, song*). Such labels are covered elsewhere.

In this document I summarise existing standards that could provide vocabulary/ontology terms to be used when annotating event sequences/segments in bioacoustic audio.

Existing standards considered

* PBcore (for audiovisual data) http://pbcore.org/

* JAMS (for music) https://jams.readthedocs.io/en/stable/jams_structure.html

* WebVTT (for subtitles) https://en.wikipedia.org/wiki/WebVTT

* Music Ontology (**MO**) http://musicontology.com/ which incorporates Timeline Ontology (**TLO**) and Event Ontology (**EO**). And related: the Segment Ontology (**SO**).

* NIST standards for evaluation of speech transcription: https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation

* Praat TextGrid http://www.fon.hum.uva.nl/praat/manual/TextGrid_file_formats.html

* Audio Value Chain Ontology (**AVCO**) / MVCO IPR Ontologies: http://ceur-ws.org/Vol-2423/DataTV2019_paper_2.pdf

* ISO BMFF aka MP4 (based on Apple QuickTime container format): https://mpeg.chiariglione.org/standards/mpeg-4/iso-base-media-file-format/overview-iso-base-media-file-format

* W3C Annotation Model https://www.w3.org/TR/annotation-model/ (WAM)

* IIIF Presentation API 3.0 https://iiif.io/api/presentation/3.0/ (PresAPI)

* DarwinCore "event" https://dwc.tdwg.org/terms/#event (dwc:Event)

Recording-level metadata that can also be tagged per-event

There are many metadata terms that are commonly associated with audio recordings, which can also be usefully tagged at the finer per-event level: latitude/longitude/altitude (as with the overall metadata, it should be made clear whether these refer to location of the event or the recording device – both can vary from moment to moment), taxon, effort, annotator. These can be reused directly from the wider AC standards.

Issues to consider

- It's common in bioacoustics to indicate events as "bounding boxes" on a spectrogram, which means events have start+end time as well as **low+high frequency**. The latter are *not* covered in most of the existing standards.
- Should there be separate terminology for contiguous nonoverlapping "segments", distinct from non-contiguous potentially-overlapping "events"? The SO deliberately implements this. I suggest that this is probably excessive for the present case.
- Should a region be given as [start time, duration] or [start time, end time]? Both are common. The TLO explicitly allows both. This seems reasonable.
- Similarly, should a start time be given as the time from the start of the recording, or as an absolute date-time value? TLO again allows for both. Absolute date-times are likely to be overly verbose for most cases, and in many datasets time-from-start is implicitly assumed. To annotate this reliably, the recording's own start date-time should ideally be logged; then, other start times can be given in seconds or h:m:s, or as integers (number of samples) on a discrete timeline where the sample rate (quantisation of time) has been specified.

Useful/interesting features from existing standards

dwc:Event allows for time instants and ranges, but mainly designed for capture events, it focusses on ranges on a resolution of days within one year. It explicitly suggests a slash-separated range syntax "[start]/[end]" which could be convenient, but seems problematic because it will need specialised parsing. Also the descriptions/examples are about capture events, not acoustic events – consider whether those should be distinguished?

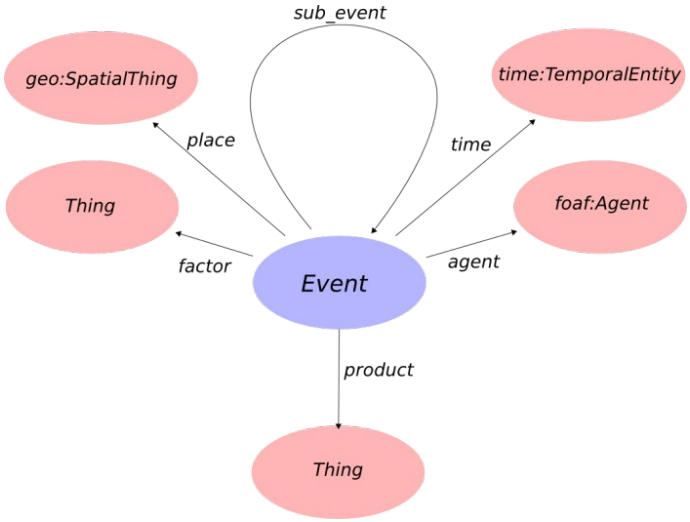
TLO allows times to be specified in various ways, e.g. as an explicit date-time or as a duration since the origin. In all cases a frame of reference needs to be defined (e.g. UTC, or temporal origin). TLO makes clear the distinction between a continuous time-line, and a discrete time-line with a specified quantisation (on which times can be specified as integers). PresAPI does differently: time is simply specified in seconds, relative to the beginning of a "canvas".

AVCO has "Segment" objects but also the concept of "Track", intended to designate e.g. when different music recordings are superimposed in a DJ mix. Metadata can be associated with a Track which can then cascade to every Segment associated with it. This is not likely to be a key demand in bioacoustics, though such a structure could be used e.g. to group sound events emitted by a particular species or individual.

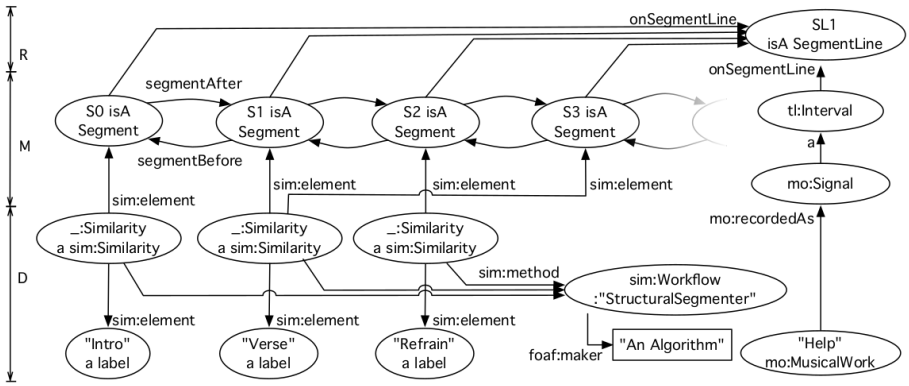
Hierarchical segmentation: the EO allows an "Event" to have "sub_event" items (see diagram), which allows for arbitrary hierarchical nesting. The SO handles this slightly differently, related to its approach with segments as contiguous regions.

SO allows for multiple levels of annotation in a very different way, by allowing multiple SegmentLines to be given, along with a SegmentLineMap that can imply containment relationships. This is perhaps cumbersome for bioacoustic annotations, and more appropriate for the music audio features that were the intended use case.

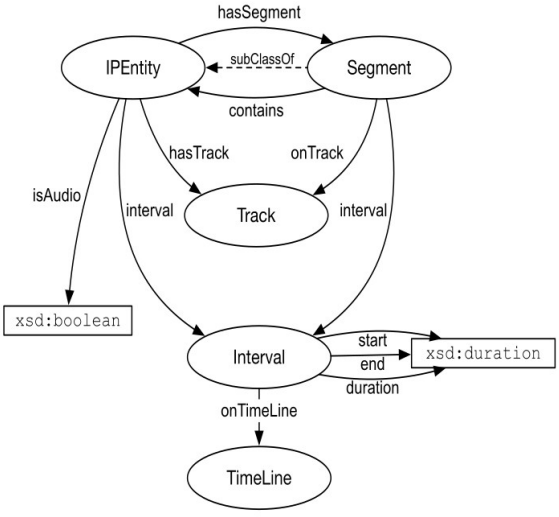
“Event” and “sub_event” from Event Ontology:



“SegmentLine” and “Segment” from Segment Ontology:



“Segment” and “Track” from Audio Value Chain Ontology:



Recommendations

Adopt the TLO and EO core concepts. “Timeline” (continuous or discrete, with a temporal origin which can be given as the tl:start in the audio data item’s overall metadata [usually in UTC]). “Event” (which can be a point-like Instant or an Interval). Interval has tl:start and either tl:duration or tl:end (user’s choice). Instant Temporal values can be given either in absolute (could use tl:beginsAtDateTime) or relative to the origin. The latter can be in seconds (tl:start), or in integer samples (could use tl:beginsAtInt) (need to specify tl:sampleRate for the overall annotation).

Do not distinguish between Event and Segment (unneeded complexity). Do not use Track either.

Allow for hierarchical structure in transcriptions, through Event’s “sub_event” approach. This approach can be transparently ignored in most applications that do not use hierarchical annotation. ***– This issue may need discussion.***

Need ***new terminology*** for upper and lower bounding boxes: **freq_low** and **freq_high**. (This terminology is also proposed in the ongoing ann-o-mate work that Ed and Dan are involved in.) These should always be specified in Hz (some applications use e.g. kHz as units – this will need to be explicit). These are optional, and it can often be reasonable to state one and not the other. They refer to properties of the sound event being annotated, in principle independent of the recording medium.