University of Zurich[UZH]

idiap RESEARCH INSTITUTE

Automatic Audio Segmentation

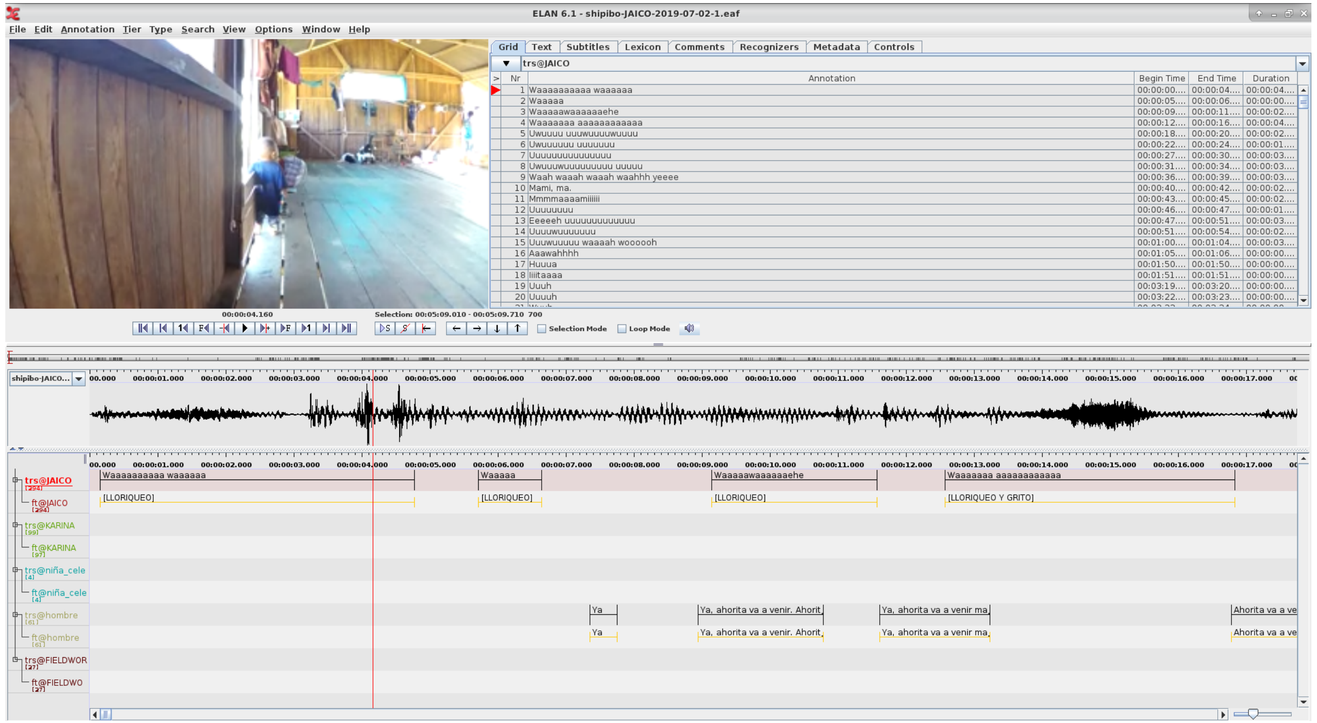
Shipibo Dataset

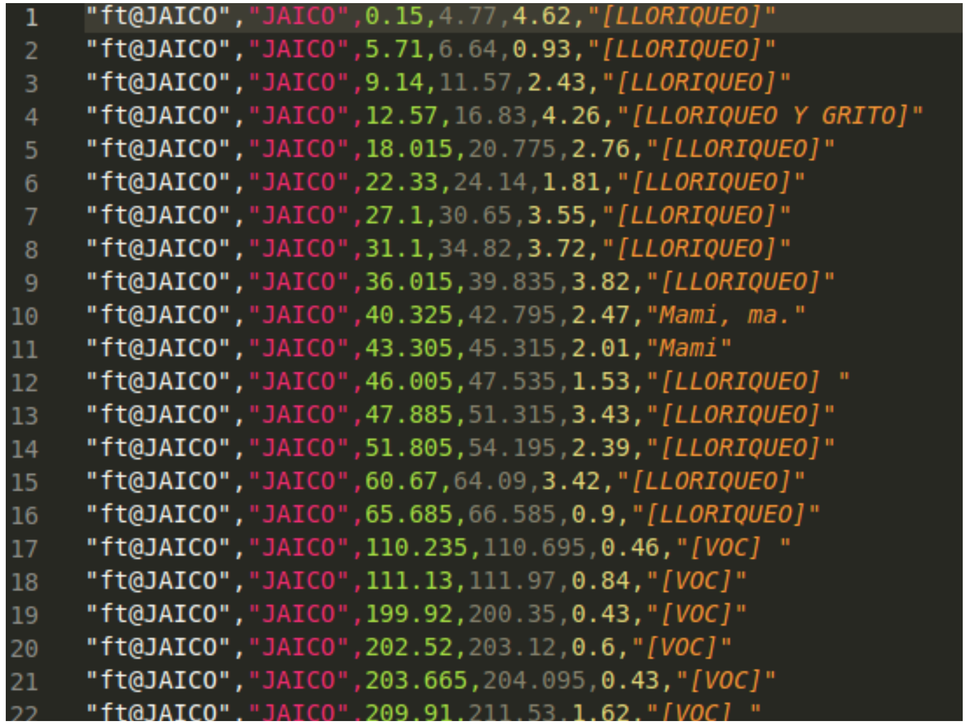
Eklavya SARKAR
July 2021

Overview

- Segmentation Pipeline
- Metric
- Models
- Tool
- File Formats
- Summary
- Future Work

Annotations



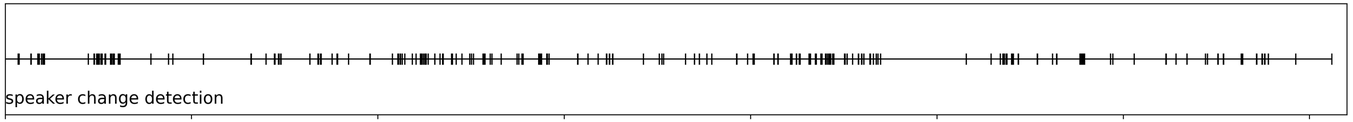


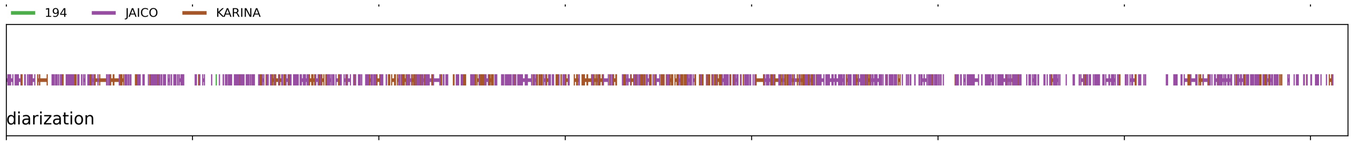
Segmentation Pipeline

PyAnnote-Audio

pyannote

1. Raw Audio
2. Feature Extraction
3. Speech Activity Detection
4. Speaker Change Detection
5. Embedding
6. Clustering
7. Re-segmentation





Metric

Diarization Error Rate (DER):

- **Speaker Error**: wrong speaker ID is assigned within a speech region.
- **False Alarm Speech**: non-speech region is incorrectly marked as containing speech.
- **Missed Speech**: speech region is incorrectly marked as not containing speech.
- Lower is better

$$\text{DER} = \frac{\text{Missed Detection} + \text{False Alarm} + \text{Speaker Confusion}}{\text{Total Speaking Time}}$$

Models

ID	Model	Train	Dev	Test	DER [%]
M1	Pre-trained	AMI[a]		JAICO[1]	164
M2		DIHARD[b]			175
M3		AMI[a]		JOHANI[2]	135
M4		DIHARD[b]			141
M5	Fine-tuned	AMI[a]	JAICO[1]	JOHANI[2]	
M6		DIHARD[b]			
M7	Trained	JAICO[1]		JOHANI[2]	

- [1]JAICO = `shipibo-JAICO-2019-07-02-1.wav` (1 hr.)
- [2]JOHANI = `shipibo-JOHANI-2019-07-13-1.wav` (18 mins.)

- [a]AMI = 70 hrs.
- [b]DIHARD = 16 hrs.

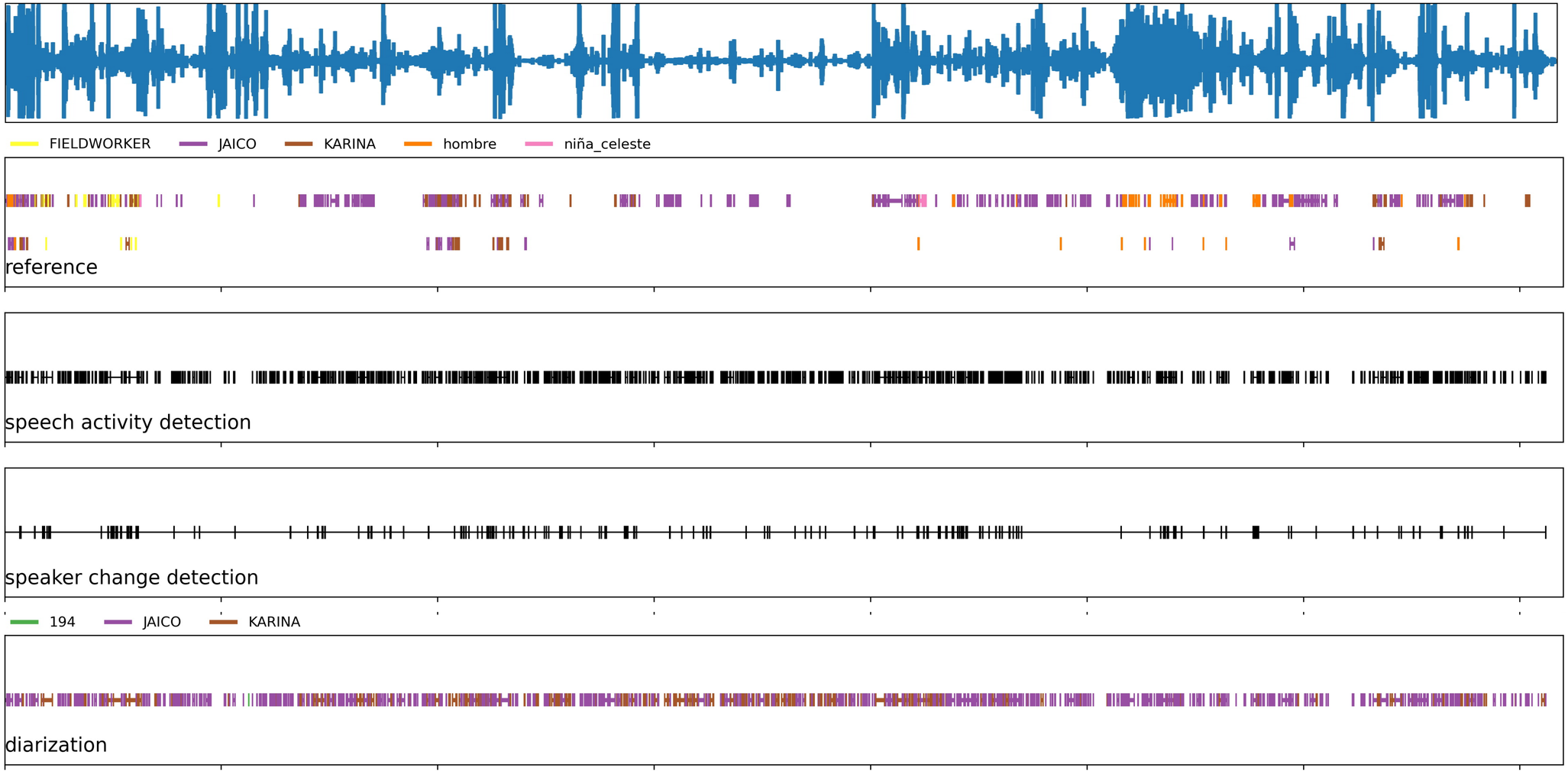
Models

ID	Model	Train	Dev	Test	Missed Detection [%]	False Alarm [%]	Speaker Confusion [%]	DER [%]
M1	Pre-trained	AMI[a]		JAICO[1]	370	854	275	164
M2		DIHARD[b]			230	964	401	175
M3		AMI[a]		JOHANI[2]	140	440	288	135
M4		DIHARD[b]			118	469	317	141
M5	Fine-tuned	AMI[a]	JAICO[1]	JOHANI[2]				
M6		DIHARD[b]						
M7	Trained	JAICO[1]		JOHANI[2]				

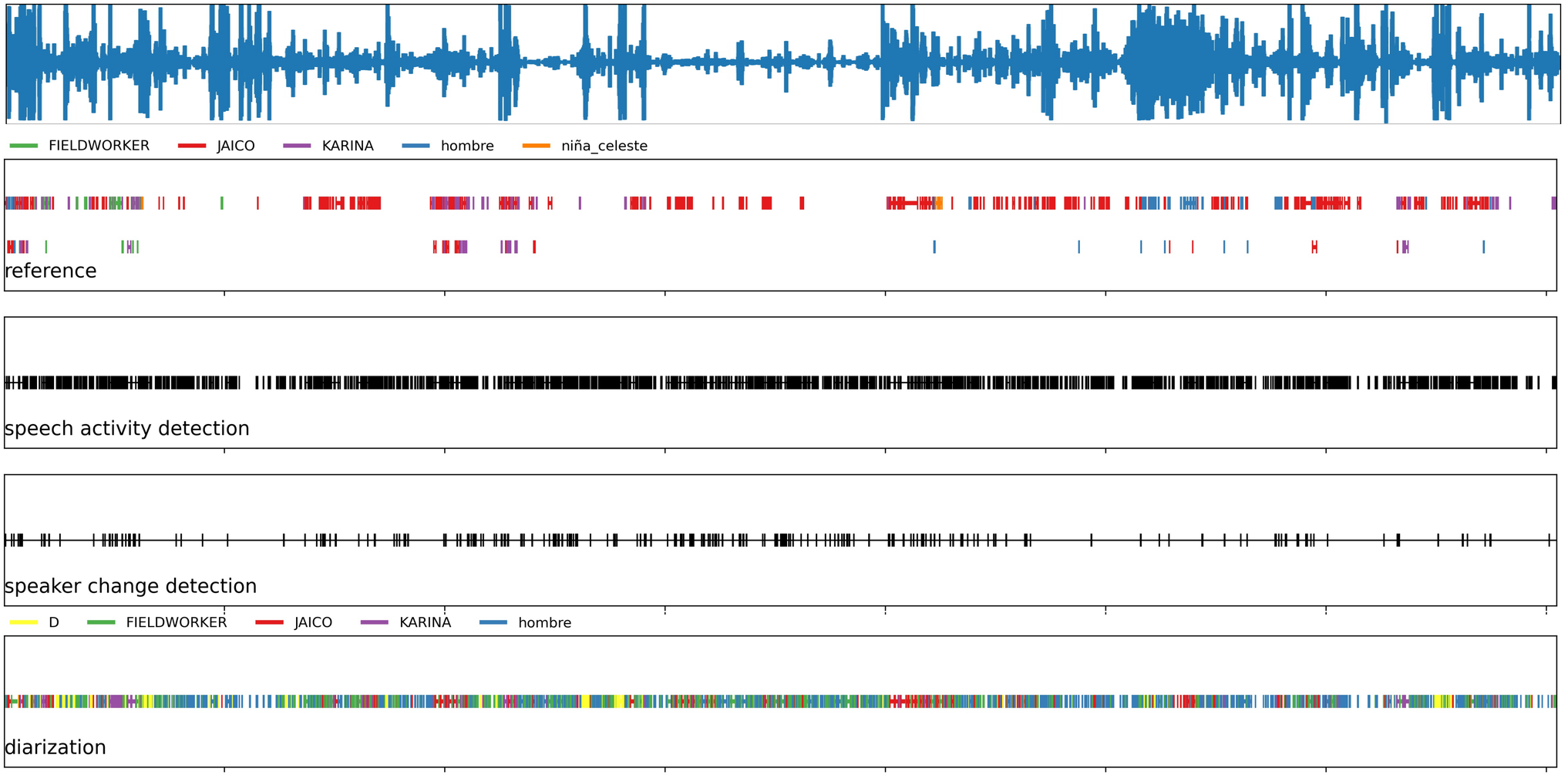
- [1]JAICO = `shipibo-JAICO-2019-07-02-1.wav` (1 hr.)
- [2]JOHANI = `shipibo-JOHANI-2019-07-13-1.wav` (18 mins.)

- [a]AMI = 70 hrs.
- [b]DIHARD = 16 hrs.

M1 Model


FIELDWORKER
JAICO
KARINA
hombre
niña_celeste
reference
speech activity detection
speaker change detection
194
JAICO
KARINA
diarization

M2 Model


FIELDWORKER
JAICO
KARINA
hombre
niña_celeste
reference
speech activity detection
speaker change detection
D
FIELDWORKER
JAICO
KARINA
hombre
diarization

Models

ID	Model	Train	Dev	Test	Missed Detection [%]	False Alarm [%]	Speaker Confusion [%]	DER [%]
M1	Pre-trained	AMI[a]		JOHANI[1]	370	854	275	164
M2		DIHARD[b]			230	964	401	175
M3		AMI[a]		JAICO[2]	140	440	288	135
M4		DIHARD[b]			118	469	317	141
M5	Fine-tuned	AMI[a]	JAICO[2]	JOHANI[1]				
M6		DIHARD[b]						
M7	Trained	JAICO[2]		JOHANI[1]				

- [1]JOHANI = `shipibo-JOHANI-2019-07-13-1.wav` (18 mins.)
- [2]JAICO = `shipibo-JAICO-2019-07-02-1.wav` (1 hr.)

- [a]AMI = 70 hrs.
- [b]DIHARD = 16 hrs.

Tool

Available at:

- https://gitlab.idiap.ch/esarkar/autodiarization
- Add members ?

```
$ conda env create -f environment.yml -n autoseg
$ conda activate autoseg
$ python pyannote.py /path/to/audio/file.wav
>>> Saving segmentation as `audio_file.rttm`.
```

Questions:

- Diarize all provided segments ? (< 26 hrs.)
- GPUs ?
- Required output format ?
  - RTTM? CSV ? Elan-readable?

File Formats

RTTM

Rich Transcription Time Marked files are space-delimited text files containing one turn per line, each containing:

- **Type**: segment type; should always by SPEAKER
- **File ID**: file name of the audio recording without format suffix (e.g., shipibo-JOHANI-2019-07-13-1)
- **Channel ID**: channel (1-indexed) that turn is on; should always be 1
- **Turn Onset**: onset of turn in seconds from beginning of recording
- **Turn Duration**: duration of turn in seconds
- **Orthography Field**: should always by <NA>
- **Speaker Type**: should always be <NA>
- **Speaker Name**: name of speaker of turn; should be unique within scope of each file
- **Confidence Score**: system confidence (probability) that information is correct; should always be <NA>
- **Signal Lookahead Time**: should always be <NA>

Eg: SPEAKER shipibo-JAICO-2019-07-02-1 1 0.15 4.62 <NA> <NA> JAICO <NA> <NA>

File Formats

UEM

Un-partitioned Evaluation Map (UEM) files are used to specify the scoring regions within each recording.

For each scoring region, the UEM file contains a line with the following four space-delimited fields:

- **File ID**: file name of the audio recording without format suffix (e.g., shipibo-JOHANI-2019-07-13-1)
- **Channel ID**: channel (1-indexed) that scoring region is on
- **Onset**: evaluation scoring segment start time [s]
- **Offset**: evaluation scoring segment end time [s]

Eg: shipibo-JAICO-2019-07-02-1 1 0.15 4.77

Summary

- Pre-trained models overly sensitive to noise (thresholds tuned on AMI / DIHARD dataset).
- To improve performance:
  - Currently fine-tuning pre-trained models on Shipibo dataset.
  - Train (from scratch) a speech activity detector on Shipibo dataset.
- Tool available on GitLab.

Future Work

Alternative Methods:

- i-vectors
- x-vectors
- VBx-HMM

Alternative Tools:

- Kaldi
- SpeechBrain

Additional outputs:

- Overlapping speech
- Other tags

Thank You