



University of
Zurich^{UZH}

EPFL

idiap
RESEARCH INSTITUTE

Automatic Audio Segmentation

Shipibo Dataset

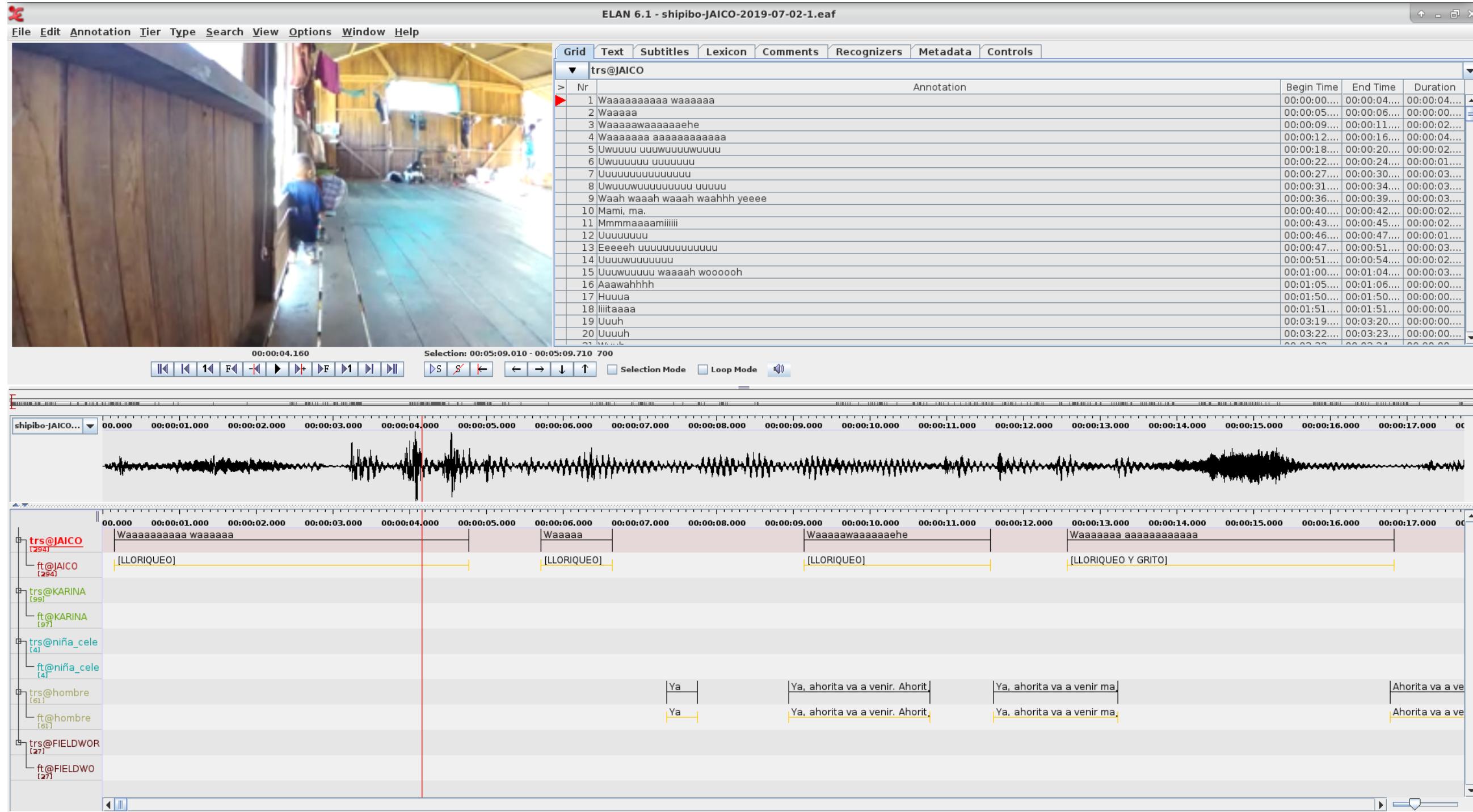
Eklavya SARKAR
July 2021

Overview

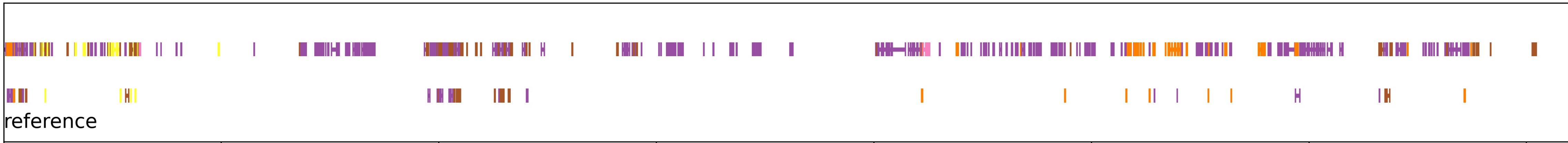


- Segmentation Pipeline
- Metric
- Models
- Tool
- File Formats
- Summary
- Future Work

Annotations



— FIELDWORKER — JAICO — KARINA — hombre — niña_celeste



```

1 "ft@JAICO", "JAICO", 0.15, 4.77, 4.62, "[LLORIQUEO]"
2 "ft@JAICO", "JAICO", 5.71, 6.64, 0.93, "[LLORIQUEO]"
3 "ft@JAICO", "JAICO", 9.14, 11.57, 2.43, "[LLORIQUEO]"
4 "ft@JAICO", "JAICO", 12.57, 16.83, 4.26, "[LLORIQUEO Y GRITO]"
5 "ft@JAICO", "JAICO", 18.015, 20.775, 2.76, "[LLORIQUEO]"
6 "ft@JAICO", "JAICO", 22.33, 24.14, 1.81, "[LLORIQUEO]"
7 "ft@JAICO", "JAICO", 27.1, 30.65, 3.55, "[LLORIQUEO]"
8 "ft@JAICO", "JAICO", 31.1, 34.82, 3.72, "[LLORIQUEO]"
9 "ft@JAICO", "JAICO", 36.015, 39.835, 3.82, "[LLORIQUEO]"
10 "ft@JAICO", "JAICO", 40.325, 42.795, 2.47, "Mami, ma."
11 "ft@JAICO", "JAICO", 43.305, 45.315, 2.01, "Mami"
12 "ft@JAICO", "JAICO", 46.005, 47.535, 1.53, "[LLORIQUEO] "
13 "ft@JAICO", "JAICO", 47.885, 51.315, 3.43, "[LLORIQUEO]"
14 "ft@JAICO", "JAICO", 51.805, 54.195, 2.39, "[LLORIQUEO]"
15 "ft@JAICO", "JAICO", 60.67, 64.09, 3.42, "[LLORIQUEO]"
16 "ft@JAICO", "JAICO", 65.685, 66.585, 0.9, "[LLORIQUEO]"
17 "ft@JAICO", "JAICO", 110.235, 110.695, 0.46, "[VOC] "
18 "ft@JAICO", "JAICO", 111.13, 111.97, 0.84, "[VOC]"
19 "ft@JAICO", "JAICO", 199.92, 200.35, 0.43, "[VOC]"
20 "ft@JAICO", "JAICO", 202.52, 203.12, 0.6, "[VOC]"
21 "ft@JAICO", "JAICO", 203.665, 204.095, 0.43, "[VOC]"
22 "ft@JAICO", "JAICO", 209.91, 211.53, 1.62, "[VOC]"

```

Segmentation Pipeline

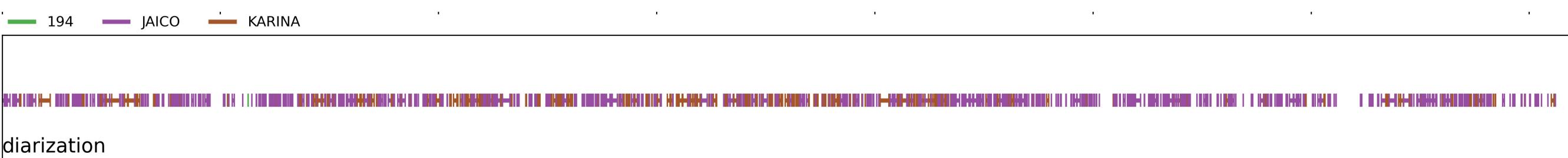
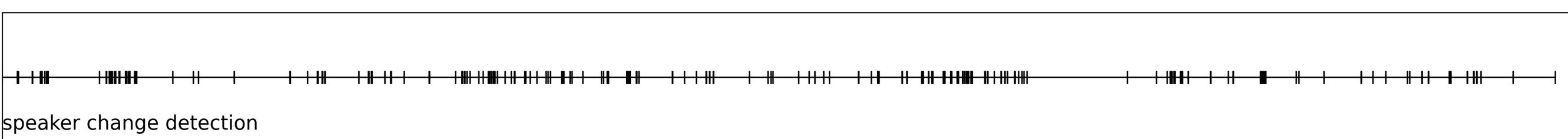
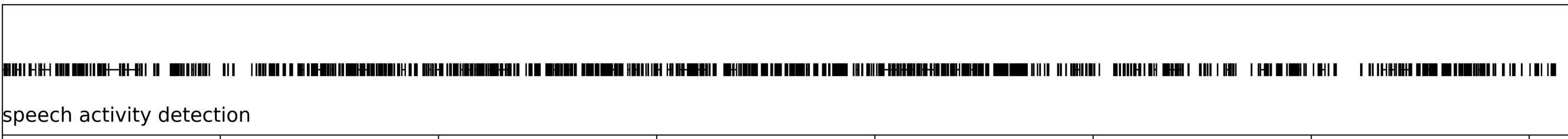
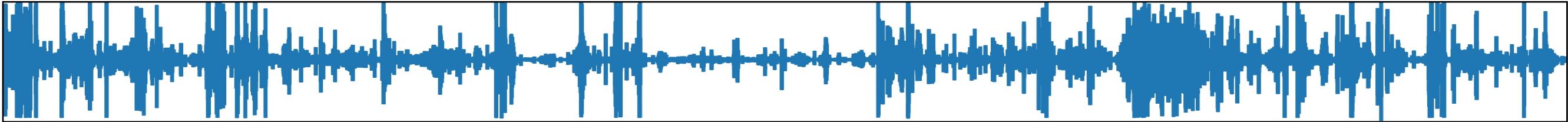


PyAnnote-Audio



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.

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

Models



ID	Model	Train	Dev	Test	DER
M1	Pre-trained	AMI		JOHANI ¹	164%
M2		DIHARD			175%
M3	Fine-tuned	AMI	JAICO ²	JOHANI ¹	
M4		DIHARD			
M5	Trained	JAICO ²		JOHANI ¹	

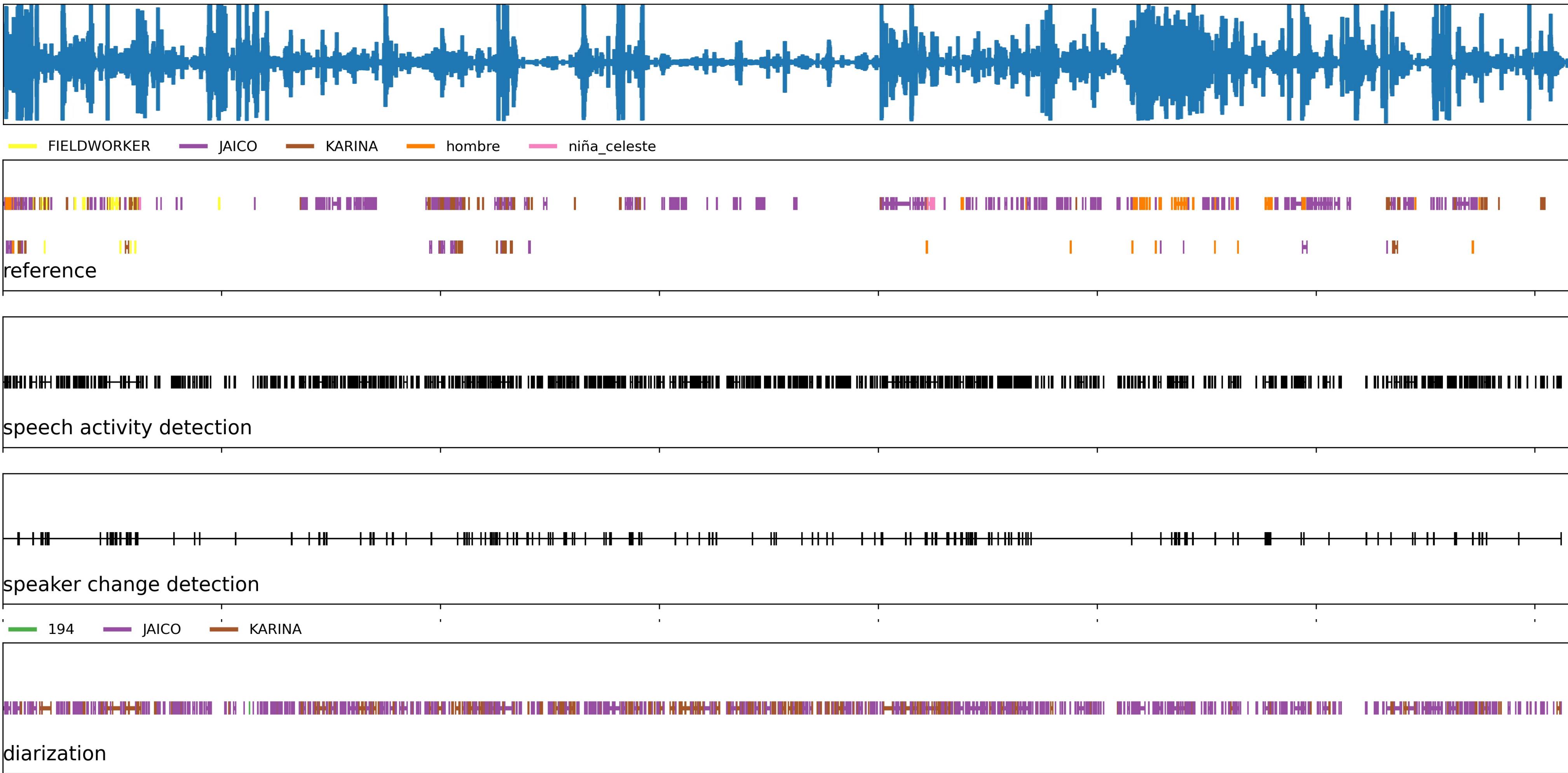
- JOHANI¹ = shipibo-JOHANI-2019-07-13-1.wav (18 mins.)
- JAICO² = shipibo-JAICO-2019-07-02-1.wav (1 hr.)

Models

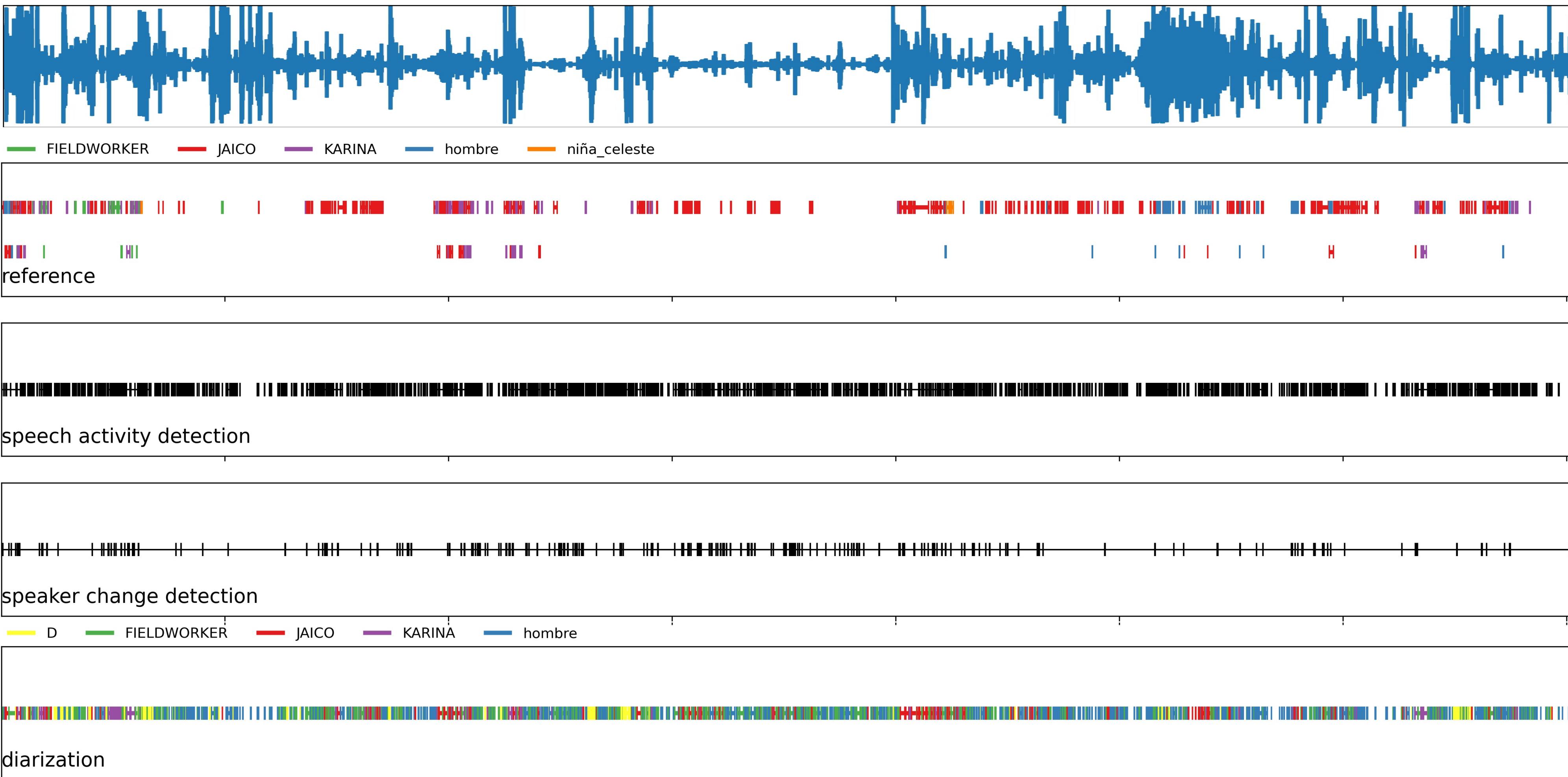
ID	Model	Train	Dev	Test	Missed Detection	False Alarm	Speaker Confusion	DER
M1	Pre-trained	AMI		JOHANI ¹	370%	854%	275%	164%
M2		DIHARD						175%
M3	Fine-tuned	AMI	JAICO ²	JOHANI ¹				
M4		DIHARD						
M5	Trained	JAICO ²		JOHANI ¹				

- JOHANI¹ = shipibo-JOHANI-2019-07-13-1.wav (18 mins.)
- JAICO² = shipibo-JAICO-2019-07-02-1.wav (1 hr.)

M1 Model



M2 Model



Models



ID	Model	Train	Dev	Test	DER
M1	Pre-trained	AMI		JOHANI ¹	164%
M2		DIHARD			175%
M3	Fine-tuned	AMI	JAICO ²	JOHANI ¹	
M4		DIHARD			
M5	Trained	JAICO ²		JOHANI ¹	

- JOHANI¹ = shipibo-JOHANI-2019-07-13-1.wav (18 mins.)
- JAICO² = shipibo-JAICO-2019-07-02-1.wav (1 hr.)

Tool



Available at:

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

Questions:

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



```
$ 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`.
```

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 be 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 be <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