MINISTERIO DE CIENCIA E INNOVACIÓN
Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia
AGENCIA ESTATAL DE INVESTIGACIÓN
HiTZ Hizkuntza Teknologiako Zentroa Basque Center for Language Technology
ahoLAB

Diarization

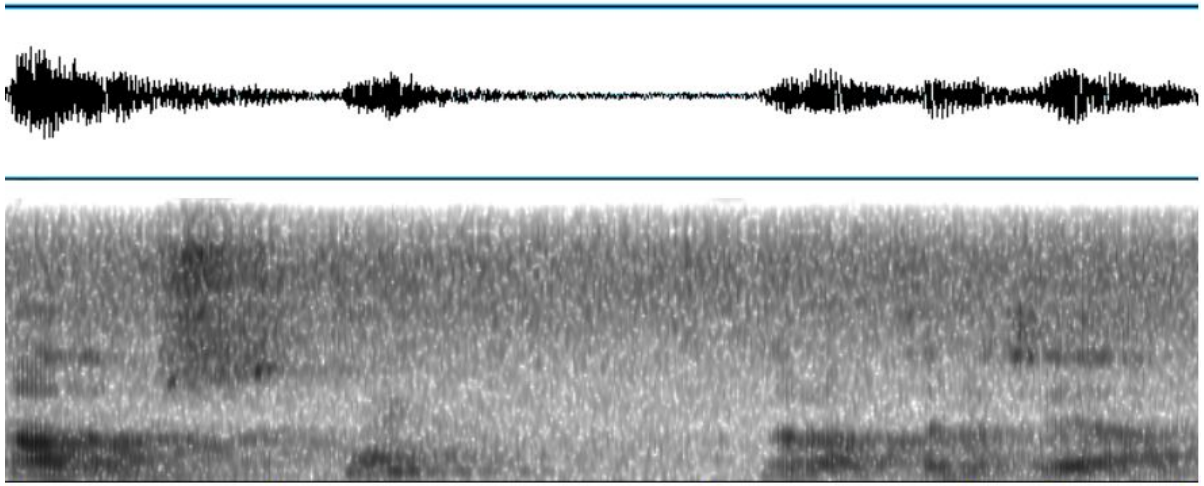
Preliminary results

Christoforos Souganidis

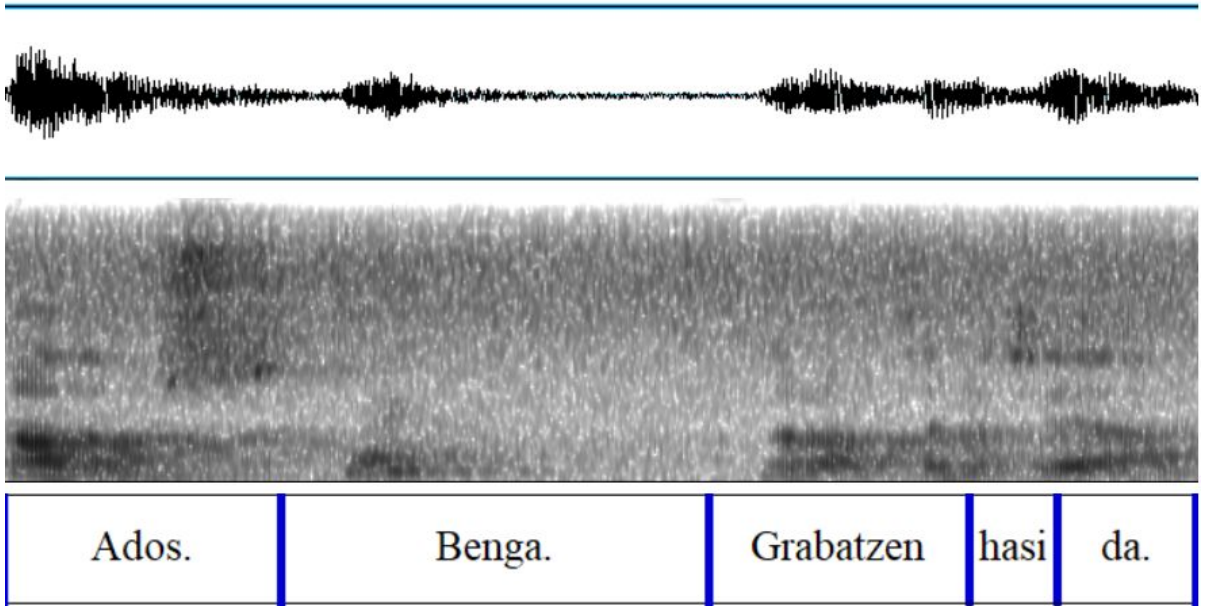
Este proyecto con referencia 2022/TL22/00215335 está financiado por el Ministerio de Transformación Digital y por el Plan de Recuperación, Transformación y Resiliencia – Financiado por la Unión Europea – NextGenerationEU.

19/02/2024

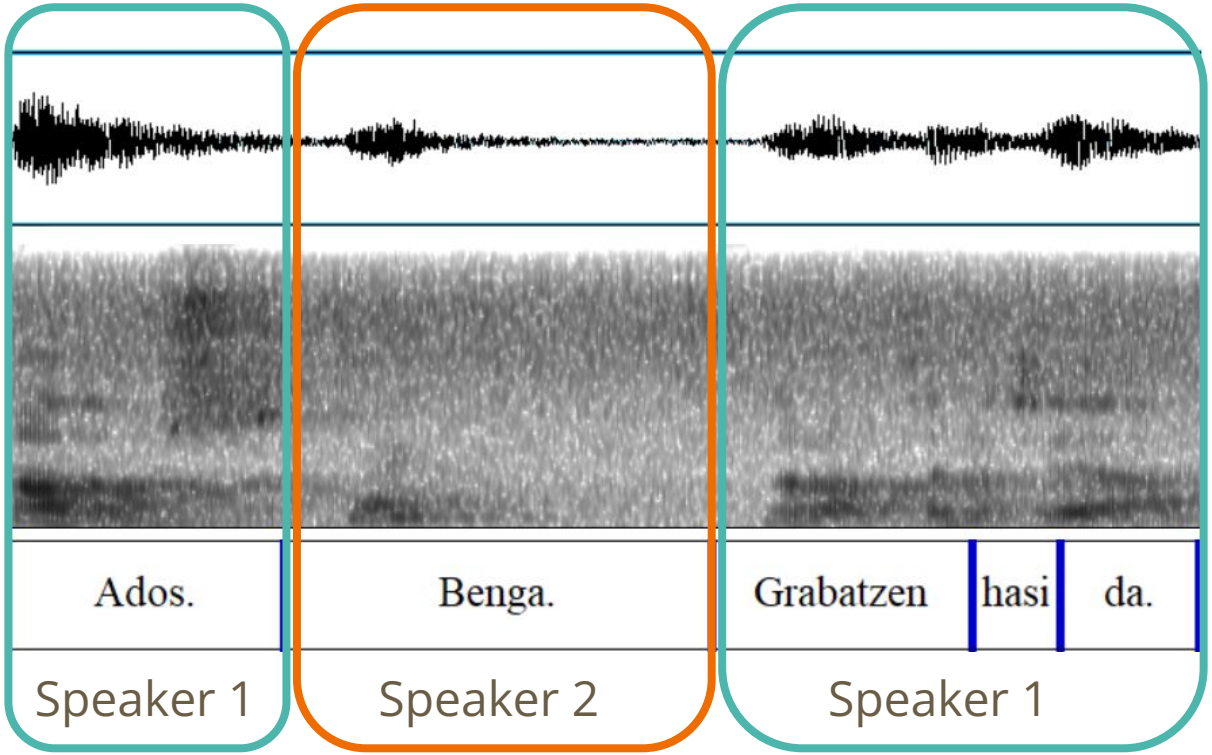
Speaker diarization



Speaker diarization



Speaker diarization



Why diarize?

- Uses of speaker diarization:
  - Automatic speech recognition (ASR)
  - Language identification
  - Speaker verification

- For ILENIA/Ikergaitu:
  - Bilingual ASR → ASR + Language Identification

- In order to diarize:
  - Voice Activity Detection (VAD): Presence/Absence of human voice

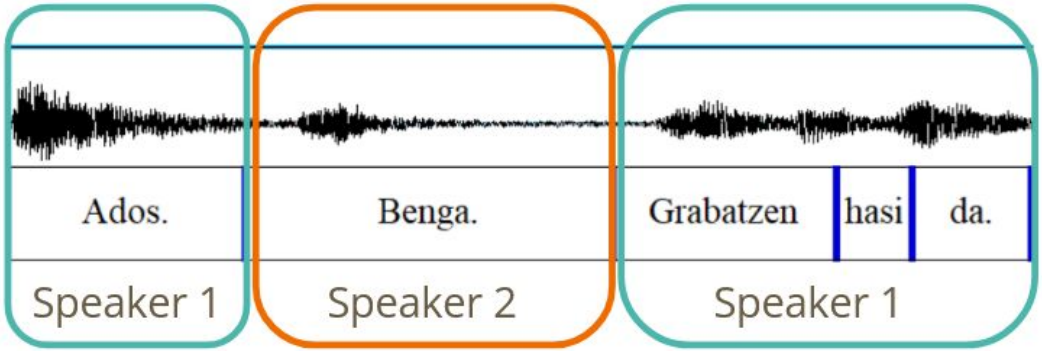
Diarization: Itzuli

- Itzuli [1]:
  - ASR-EUS
  - Speaker diarization

```
<segment speaker="38" language="" start="3.58" end="4.06">
    <word start="3.58" end="4.06" conf="0.85">Ados.</word>
</segment>
<segment speaker="48" language="" start="4.80" end="6.50">
    <word start="4.80" end="5.25" conf="1.00">Grabatzen</word>
    <word start="5.25" end="5.40" conf="0.91">hasi</word>
    <word start="5.40" end="5.64" conf="0.61">dira.</word>
</segment>
```

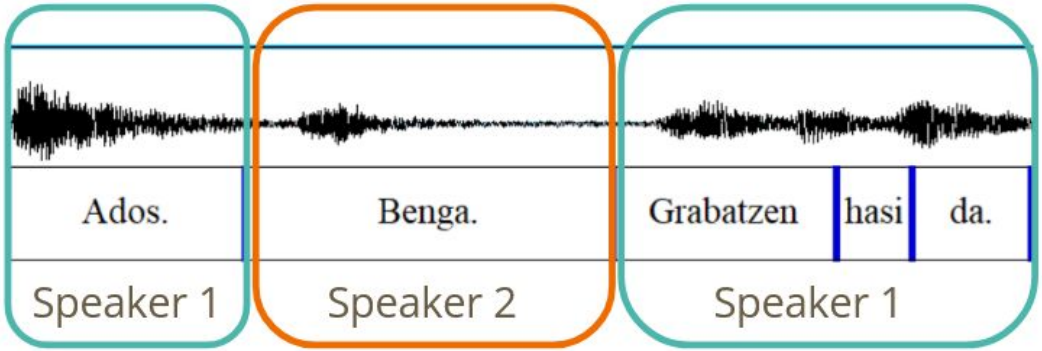
[1] https://itzuli.vicomtech.org

Diarization: Itzuli




```
<segment speaker="38" language="" start="3.58" end="4.06">
    <word start="3.58" end="4.06" conf="0.85">Ados.</word>
</segment>
<segment speaker="48" language="" start="4.80" end="6.50">
    <word start="4.80" end="5.25" conf="1.00">Grabatzen</word>
    <word start="5.25" end="5.40" conf="0.91">hasi</word>
    <word start="5.40" end="5.64" conf="0.61">dira.</word>
</segment>
```

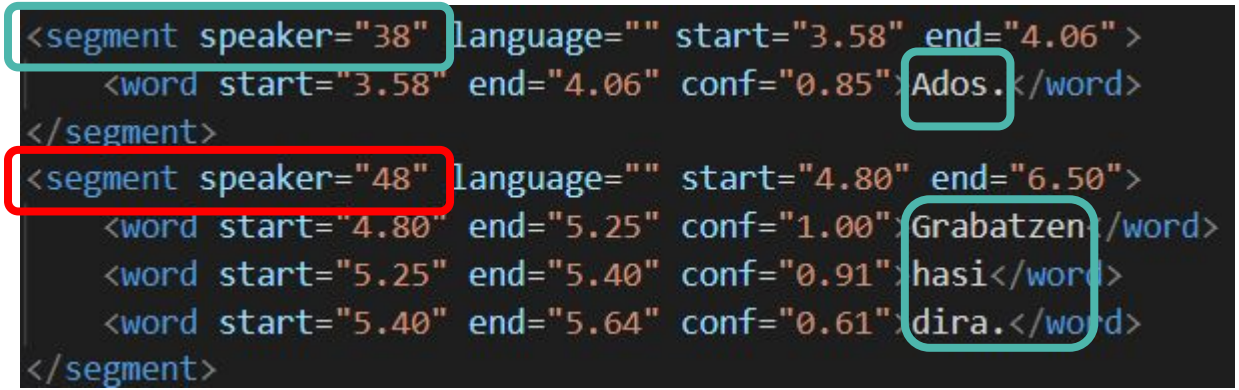
Diarization evaluation metrics



Missed detection

Diarization evaluation metrics

Speaker confusion



<word start="3.58" end="4.06" conf="0.85">Ados.</word>

<word start="4.80" end="5.25" conf="1.00">Grabatzen</word>
<word start="5.25" end="5.40" conf="0.91">hasi</word>
<word start="5.40" end="5.64" conf="0.61">dira.</word>

Diarization evaluation metrics

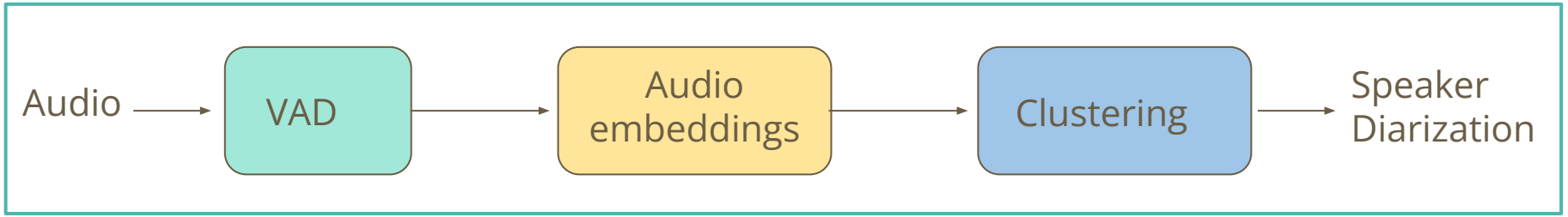
- Diarization Error Rate (DER)

$$\text{DER} = \frac{\text{Missed detection} + \text{Speaker confusion} + \text{False alarm}}{\text{Total duration}}$$

Diarization: pyannote

- Pyannote [1]:
  - Open Source & Free
  - State-of-the-art performance
  - Good overlapping speaker detection
  - Pretrained models
  - Number of speakers: Automatic/Adjustable

- Pipeline:



[1] Bredin, Hervé, et al., 2020

Databases

1. Faktoria - Euskadi Irratia (Radio program)
   - 150 h - EU
   - To be annotated, public call TBA
2. Conquis & En Jake - EITB2 (TV programs)
   - 23 h - ES
   - El Conquis, En jake (available since14/02/2024)
3. RTVE2022 DB (TV programs)
   - 25 h - ES
   - Used in the IberSpeech Challenges on Speech Technologies [1]
4. HABE (Oral exams)
   - (10493h) - EU
   - Not annotated

[1] Lleida, et al., 2023

HABE database

- Available from HABE:

CEFR LEVEL	File duration	Number of files
B1	15 min	5261
B2	15 min	15409
C1	20 min	15001
C2	10 min	1956

HABE C1 database

- Data: 448 audio files:
  - Oral exams for C1 level (HABE), between 2014-2022
  - 104 h - EU
  - 2 students (same qualification:PASS/FAIL) & 1-2 examiners

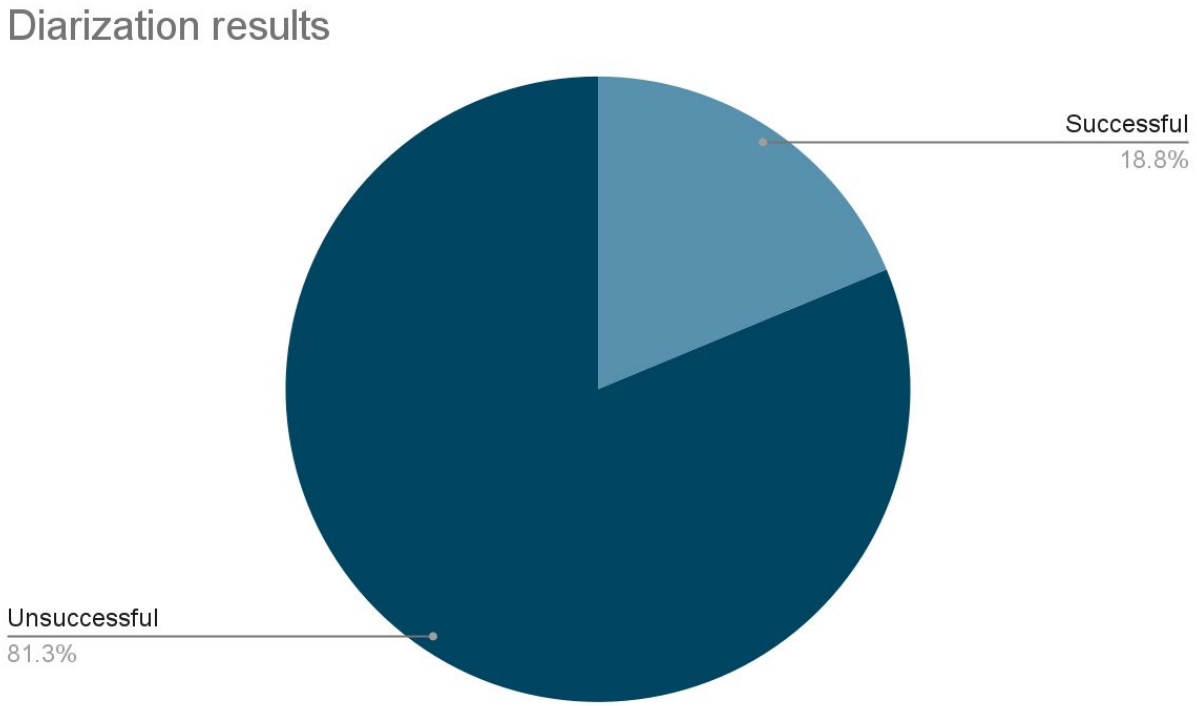
- Technical information:
  - Variable formats: Mainly in .wav or .wma, but also in .m4a or .mp3
  - Variable sampling rate: 44,1 kHZ or 16 kHz

- Sample of 16 audio files:
  - 8 PASS & 8 FAIL

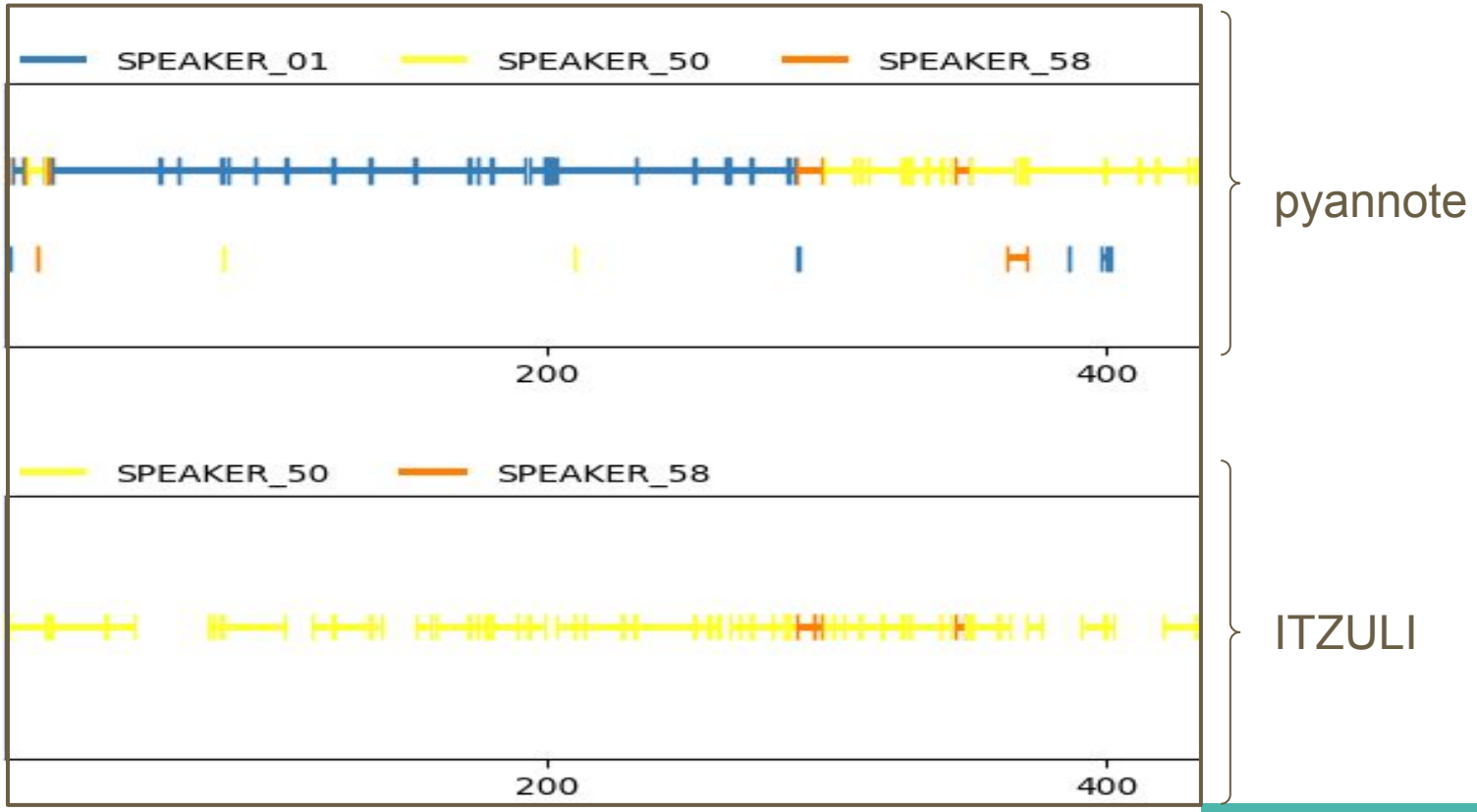
Sample files

File PASS	Student gender	File FAIL	Student gender
1413-0220-0227.wma	2 W	1413-0135-0136.wma	2 W
1513-2425-2356.wma	W + M	1513-2193-2605.wma	W + M
1613-0111-0088.wma	2 W	1613-0115-0107.wma	W + M
1723-1215-1222.wma	2 M	1713-0560-0136.wma	2 M
1823-3064-3002.wma	2 W	1823-6067-6068.wma	W + M
1923-3753-3774.wma	W + M	1913-2211-2274.wma	W + M
2113-30009-30121.wma	2 M	2113-10488-10477.mp3	2 M
2213-30368-30153.wav	2 M	2213-30615-31163.wav	W + M
	8M+8W		7W+9M

Quick analysis with Itzuli



Diarization example



Diarization subsample

- From the subsample:
  - No ground truth—> manual labeling
  - Files were also automatically diarized using ITZULI

File PASS	Student gender	Examiner gender	File FAIL	Student gender	Examiner gender
1613-0115-0107_p00.wav	W + M	M	1413-0220-0227_p00.wav	2 W	M
1713-0560-0136_p00.wav	2 M	M	1513-2425-2356_p00.wav	W + M	W
	3M+W			M+3W	

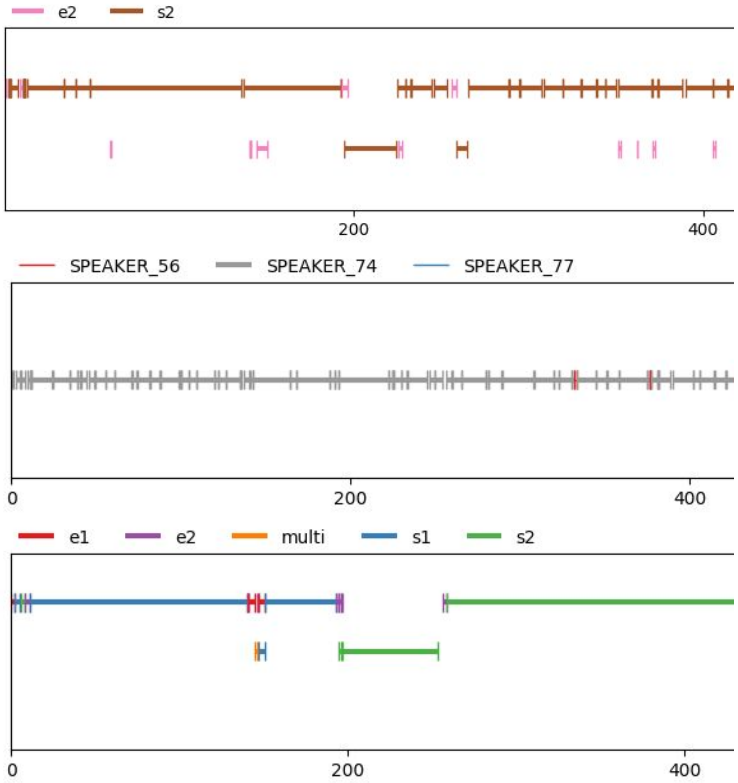
Subsample diarization results

Qualification	File	Speaker gender	Test	DER (%)	
				With overlap	Without overlap
PASS	1613-0115-0107_p00 .wav	2 W + 1 M	pyannote	**6.3**	**5.9**
			ITZULI	18	17.2
	1713-0560-0136_p00 .wav	3 M	pyannote	**12.8**	**12.6**
			ITZULI	20.3	19.8
FAIL	1413-0220-0227_p00 .wav	2 W + 1 M	pyannote	**51.1**	51.3
			ITZULI	51.4	**51.1**
	1513-2425-2356_p00 .wav	2 W + 1 M	pyannote	**14.5**	**13.2**
			ITZULI	20.2	18.1

Subsample diarization results

Qualification	File	Speaker gender	Test	DER (%)	
				With overlap	Without overlap
PASS	1613-0115-0107_p00 .wav	2 W + 1 M	pyannote	**6.3**	**5.9**
			ITZULI	18	17.2
	1713-0560-0136_p00 .wav	3 M	pyannote	**12.8**	**12.6**
			ITZULI	20.3	19.8
FAIL	1413-0220-0227_p00 .wav	2 W + 1 M	pyannote	**51.1**	51.3
			ITZULI	51.4	**51.1**
	1513-2425-2356_p00 .wav	2 W + 1 M	pyannote	**14.5**	**13.2**
			ITZULI	20.2	18.1

Test case: pyannote/Itzuli/manual



pyannote

Itzuli

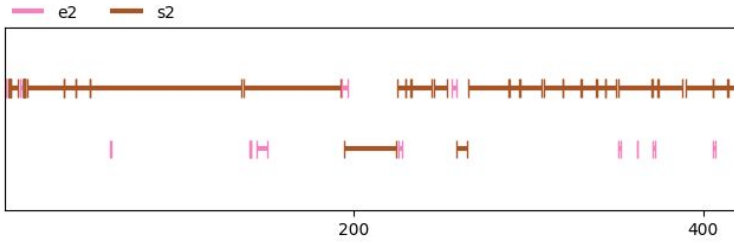
manual

Subsample diarization results

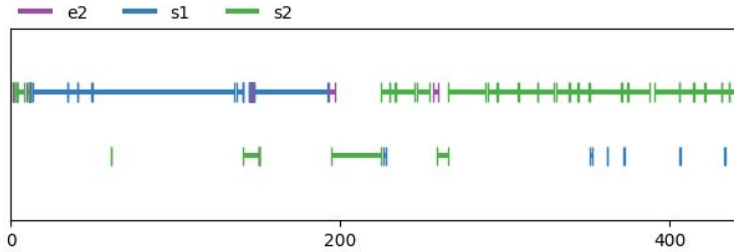
- Both pyannote and Itzuli failed to separate between the two W speakers
- pyannote can be optimised by introducing the min and max number of speakers in a file
- In pyannote-opt: min_speakers = 3 & max_speakers = 4

Qualification	File	Speaker gender	Test	DER (%)	
				With overlap	Without overlap
FAIL	1413-0220-0227_p00.wav	2 W + 1 M	pyannote	51.1	51.3
			pyannote-opt	**18.4**	**19.9**
			ITZULI	51.4	51.1

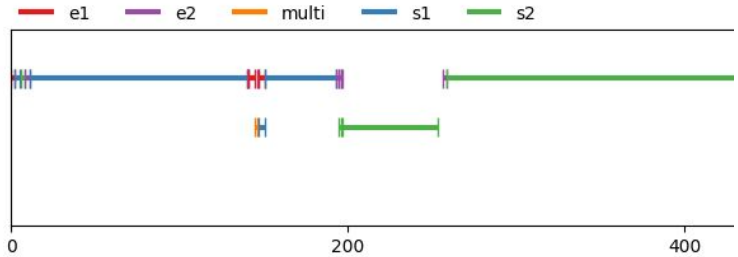
Test case: pyannote/pyannote-opt/manual



pyannote

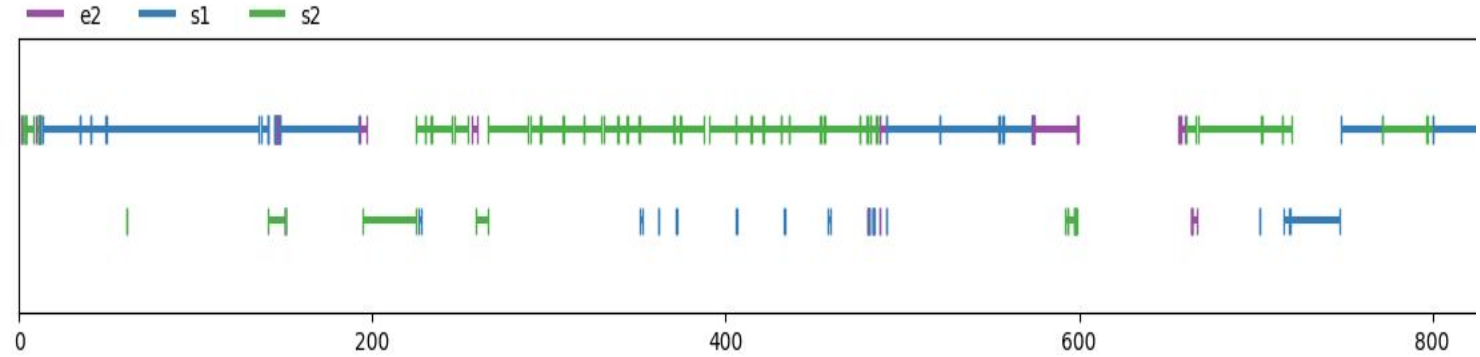


pyannote-opt

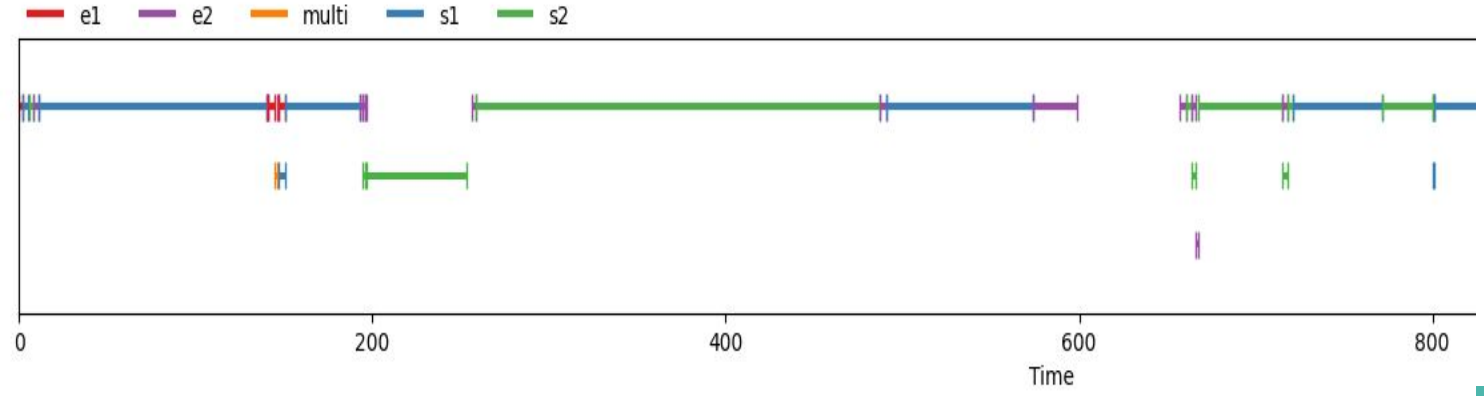


manual

Test case: pyannote-opt/manual



pyannote-opt



manual

Summary

	pyannote	Itzuli
Open source	✓	☒
Free	✓	☒
Adjustable	✓	☒
Overlap detection	☐	☒/☐
Performance	☐	☐