ALBAYZIN EVALUATIONS

Iberspeech-RTVE-Challenge 2022

Supported by

Spanish Thematic Network on Speech Technology (RTTH)

Cátedra RTVE de la Universidad de Zaragoza

Organized by

ViVoLab from Universidad de Zaragoza

RTVE

Universidad San Pablo-CEU

AuDIaS from Universidad Autónoma de Madrid

Universidad del País Vasco

http://catedrartve.unizar.es/albayzin2022.html

Four Challenges:

- ✓ Speech to Text (S2T)

  Automatic transcription of TV shows.

- ✓ Speaker Diarization and Identity Assignment (SDIA)

  Segmenting broadcast audio documents according to different speakers, linking those segments which originate from the same speaker and identify a closed set of speakers.

- ✓ Text and Speech Alignment (TaSA)
  - Synchronizing the broadcast subtitles created by respeaking
  - Aligning text and audio extracted from a plenary session of the Basque Parliament.

- ✓ Search on Speech (SoS)

  Searching in audio content a list of terms/queries.

Participation:

11 teams request the database license to participate

7 Spanish teams

TID, GTTS-EHU, VICOMTECH-UPM, AUDIAS-UAM, VIVOLAB-UZ, ECA-SIMM, BV

4 International teams

PYANNOTE (IRIT, France), TEAMIV (Intelligent Voice, UK), SPLab-IITM (India), BUT (Czech Republic)

Final participation: 7 teams

S2T	SD	TaSA	SoS
4	2+1	2	0

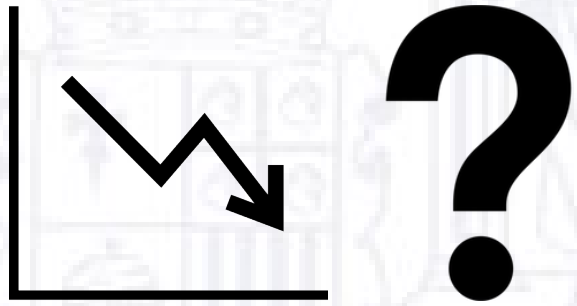
1 team has participated in two challenges

A total of 20 systems evaluated

Previous evaluations:

2018: 16 teams (85 systems evaluated)

2020: 12 teams (31 systems evaluated)



https://catedrartve.unizar.es/albayzin2022results.html

RTVE2022 Database Description

Vivolab
Aragon Institute for Engineering Resarch (I3A) Universidad de Zaragoza

Virginia Bazán , Carmen Pérez , Alberto de Prada
Corporación Radiotelevisión Española

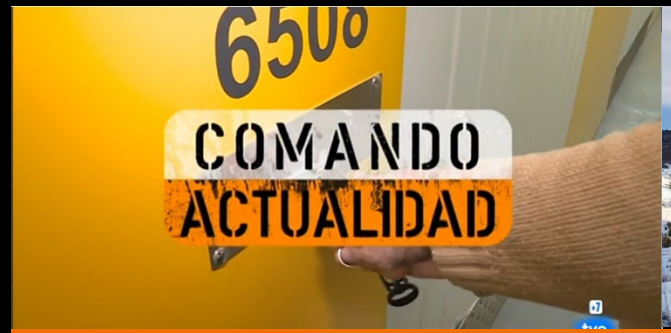
https://catedrartve.unizar.es/rtvedatabase.html

Cátedra RTVE de la Universidad de Zaragoza

July 10th, 2017
The “Corporación Radiotelevisión Española” (RTVE) and the “Universidad de Zaragoza” (UZ) signed an agreement to develop the “Cátedra RTVE de la Universidad de Zaragoza”

The purpose is to boost the technologies associated with the generation of audiovisual metadata.

One of the objectives of the Chair is to launch a set of technological challenges and provide the necessary data to test the technologies.



Databases Description: RTVE2018, RTVE2020 & RTVE2022

Video and audio quality

- ✓Video files (.mp4): multimodal diarization (2018 & 2020)
  - Format:
    - Default format in the RTVE internet channel “RTVE a la carta”
    - Video:
      - h264 standard codec with pixel format yuv420p, 1024x576 [SAR 1:1,DAR 16:9], 25 fps
      - 1500 kb/s average bit rate
    - Audio:
      - Mpeg Low-Complexity (AAC-LC) codec, sampling frequency 44100Hz stereo
      - Variable Bit rate between 48 y 96 kb/s

- ✓Audio files(.aac)
  - Format:
    - Mpeg Low-Complexity (AAC-LC) codec, sampling frequency 44100Hz stereo
    - Variable Bit rate between 48 y 96 kb/s
    - 2022: some multichannels

Database Content: RTVE2018

http://catedrartve.unizar.es/rtvedatabase.html

Training	Hours	Dev1	Hours	Task	Dev2	Hours	Task	Test	Hours	Task
20H	32:22:37	20H	9:13:13	T2S						
Agrosfera	37:34:32									
Al filo de lo imposible	6:59:54							Al filo de lo imposible	4:10:03	T2S
Asuntos publicos	61:27:00	Asuntos Públicos	8:11:00	T2S						
Arranca en Verde	4:37:35							Arranca en Verde	1:00:30	T2S
Comando actualidad	9:10:28	Comando Actualidad	7:53:13	T2S						
Dicho y Hecho	8:18:00							Dicho y Hecho	1:48:00	T2S
España en comunidad	4:53:27							España en Comunidad	8:09:32	T2S, Diarization, SoS
La mañana	218:12:00	La Mañana	1:30:00	T2S						
La tarde en 24H Economia	4:10:54									
La tarde en 24H Tertulia	17:49:40							La Tarde en 24H Tertulia	8:52:20	T2S,Diarization,Face
La tarde en 24H Entrevista	4:54:03									
La tarde en 24H El tiempo	2:20:12									
Latinoamerica en 24H	12:12:03							Latinoamerica en 24H	4:06:57	T2S, Diarization
Millennium	9:33:01				Millennium	7:42:44	Diarization, T2S, SoS	Millennium	1:52:50	T2S, SoS
Saber y Ganar	26:05:17							Saber y Ganar	2:54:53	T2S
		La noche en 24H	25:44:25	T2S	La noche en 24H	7:26:41	Diarization, T2S, SoS, Face			
	460:40:43		**52:31:51**			**15:09:25**			**41:00:05**	

Human revised manual transcriptions with speakers turns: 108 hours
Aligned transcriptions: 38 hours

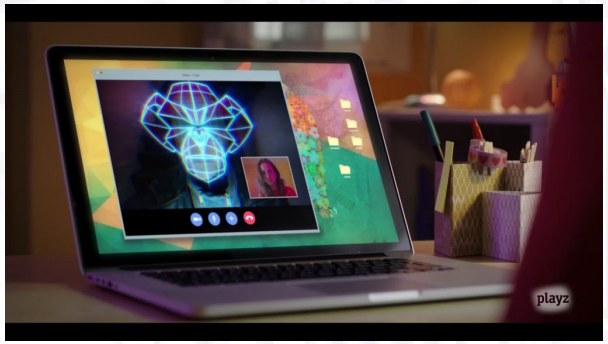
- ✓ Text associated to subtitles (24H channel, year 2017)
  - ✓ More than 3M sentences, 56 M words
  - ✓ Vocabulary of more than 160K unique words

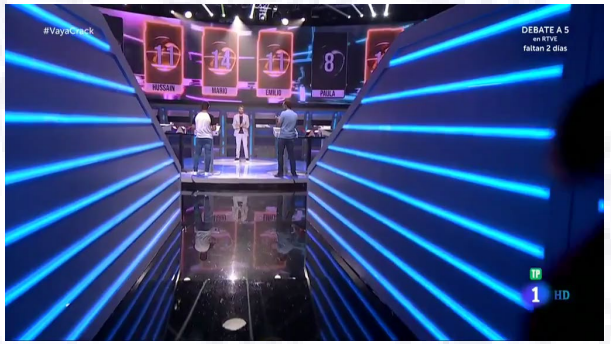
Database Content: RTVE2020

http://catedrartve.unizar.es/rtvedatabase.html

RTVE2020				
Show	Hours	S2T	SD-IA	MD-SD
Millennium	1:56:11	1:56:11	1:56:11	1:56:11
Los desayuno de tve	10:58:34	10:58:34	10:58:34	10:58:34
Comando actualidad	4:01:31	4:01:31	4:01:31	4:01:31
Ese programa del que usted me habla	1:58:36	1:58:36	1:58:36	1:58:36
Neverfilms	0:11:41	0:11:41	0:11:41	0:11:41
Si fueras tu	0:51:14	0:51:14	0:51:14	0:51:14
Bajo la red	0:59:01	0:59:01	0:59:01	0:59:01
Boca norte	1:00:46	1:00:46	1:00:46	1:00:46
Wake-up	0:57:28	0:57:28	0:57:28	0:57:28
Aquí la tierra	10:26:02	10:26:02	10:26:02	10:26:02
Versión española	2:29:12	2:29:12		
Mercado central	8:39:47	8:39:47		
Vaya crack	5:06:00	5:06:00		
Como nos reíamos	2:51:42	2:51:42		
Imprescindibles (2 canales)	3:12:31	3:12:31		
	55:40:16	**55:40:16**	**33:21:04**	**33:21:04**

161 characters have been labelled





Database Content: RTVE2022

RTVE2022				
Show	Hours	S2T	SD	TaSA
3x4	2:58:17	2:58:17	2:58:17	
A pedir de boca	3:41:38	3:41:38		
Agrosfera	4:15:20	4:15:20	4:15:20	6:38:04
Aquí la Tierra	2:46:44	2:46:44	2:46:44	2:52:35
Ateneo	1:40:09	1:40:09		
Cerámica Popular Española	1:02:35	1:02:35		
Comando Actualidad	3:59:29	3:59:29	3:59:29	
Conversatorios en Casa América	1:58:44	1:58:44		
Corazón	3:00:17	3:00:17	3:00:17	4:33:49
El cazador	3:48:22	3:48:22		
Encuestas con ruido ambiente	2:08:13	2:08:13		
Entrevistas en bruto	3:54:57	3:54:57		
España Directo	4:05:57	4:05:57	4:05:57	
Fiction (Grasa, Yrreal, Riders)	3:53:22	3:53:22	3:53:22	
Informativos UMATIC	0:59:49	0:59:49		
Jara y Sedal	2:29:17	2:29:17		
Noticias Nacional	2:14:32	2:14:32		
Saber y Ganar	4:28:28	4:28:28		
Toros	0:49:57	0:49:57		
21 different shows	**54:16:07**	**54:16:07**	**24:59:26**	**14:04:28**

74 characters have been labelled

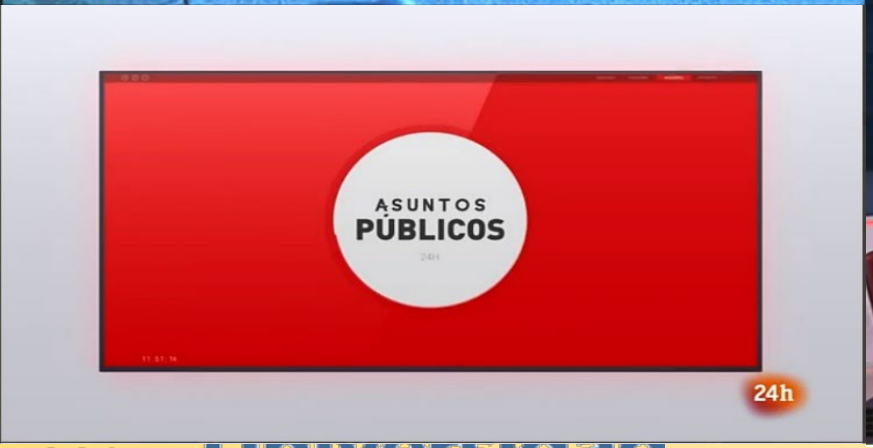
A pedir de boca

Entrevistas en bruto

Encuestas con ruido ambiente

Fiction (Grasa)

Speech to Text Challenge



Evaluation plan: (http://catedrartve.unizar.es/albayzin2022.html)

Objective:

Evaluate the state of the art in automatic speech recognition for broadcast speech transcription.

Training conditions:

✓ Open condition

Participants are free to use data to train their systems provided that these data are fully documented in the systems description paper.

Test database

54 hours of 21 TV shows covering different genres as live recordings, contests, soap opera, raw material, news, old formats recordings, …

Primary metric:

WER (Word Error Rate)

$$WER(\%) = \frac{\#I + \#D + \#S}{N} 100$$

where

$\#I$ number of insertions, $\#D$ number of deletions, $\#S$ number of substitutions and N total number of words to be recognized

Automatic speech recognition systems

Architectures:

- E2E architectures: Wav2vec2.0 based systems, RNN-Transducer,
- Hybrid CNN-TDNN-HMM
- Whisper

Acoustic corpus

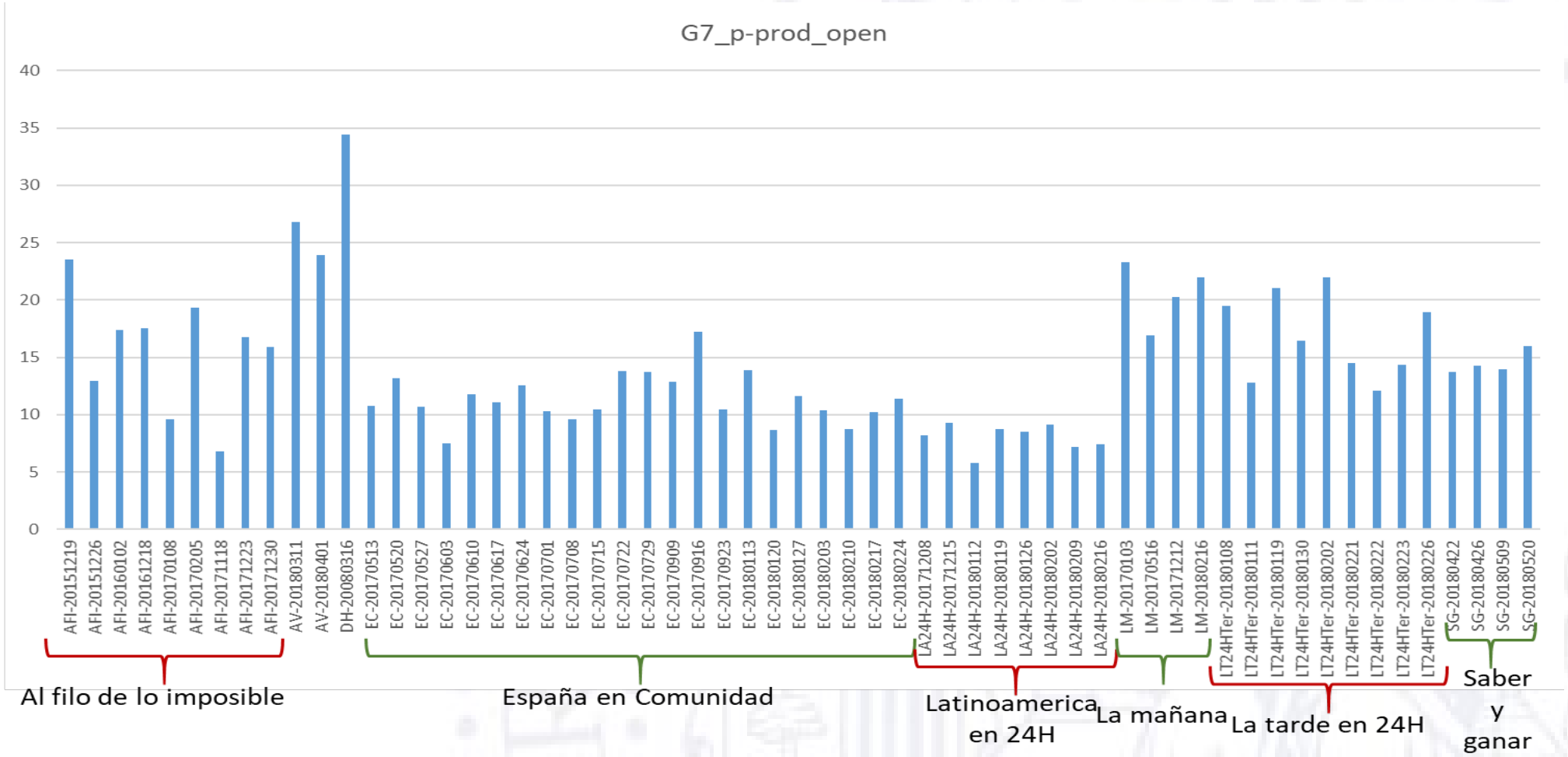
- RTVE2018/2020 databases
- Spanish CommonVoice

Text corpus

- RTVE databases transcriptions + subtitles from different sources

Previous challenges results

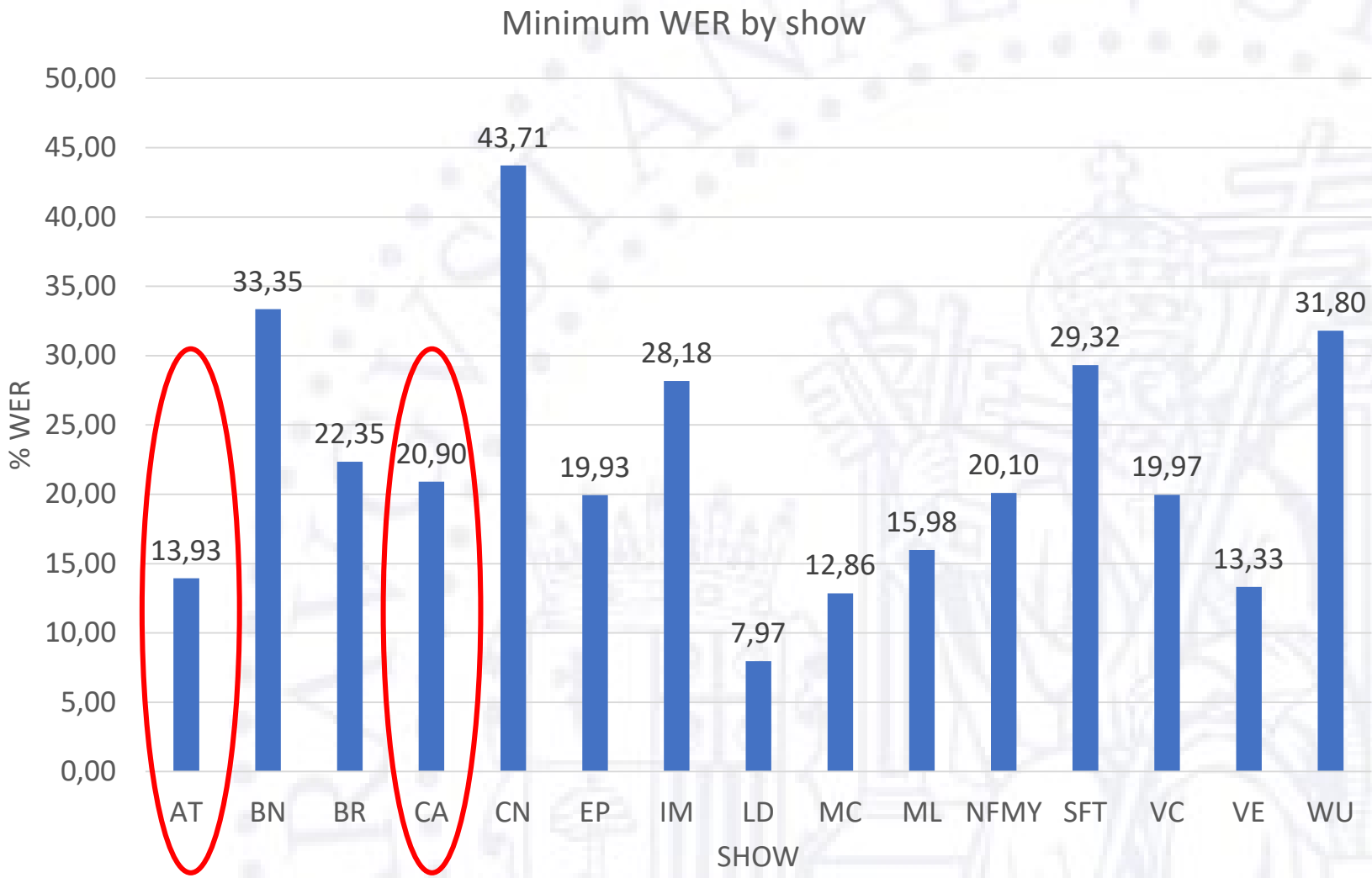
2018:
WER 16,45



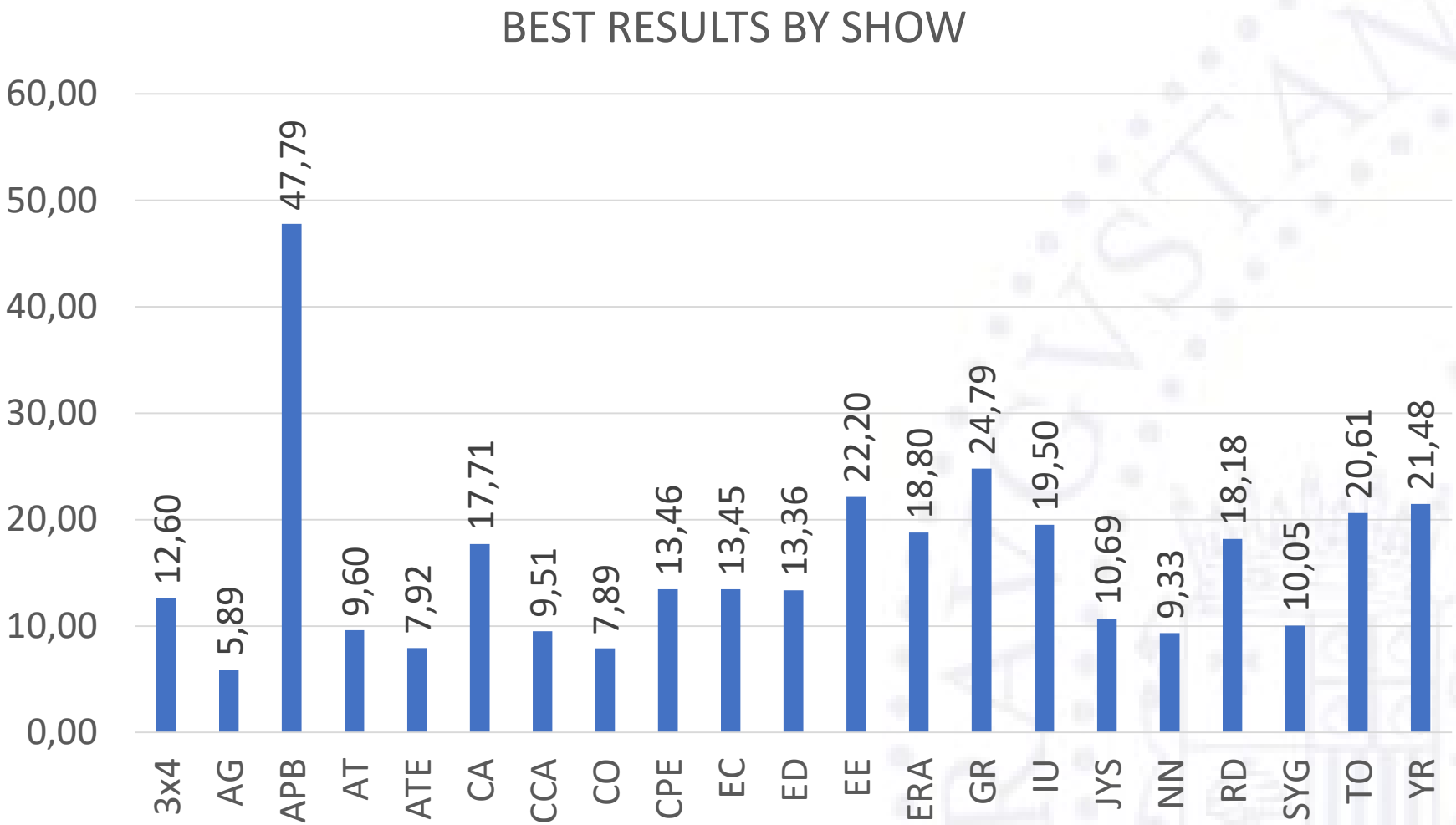
Previous challenges results

2020:
WER 16,04%

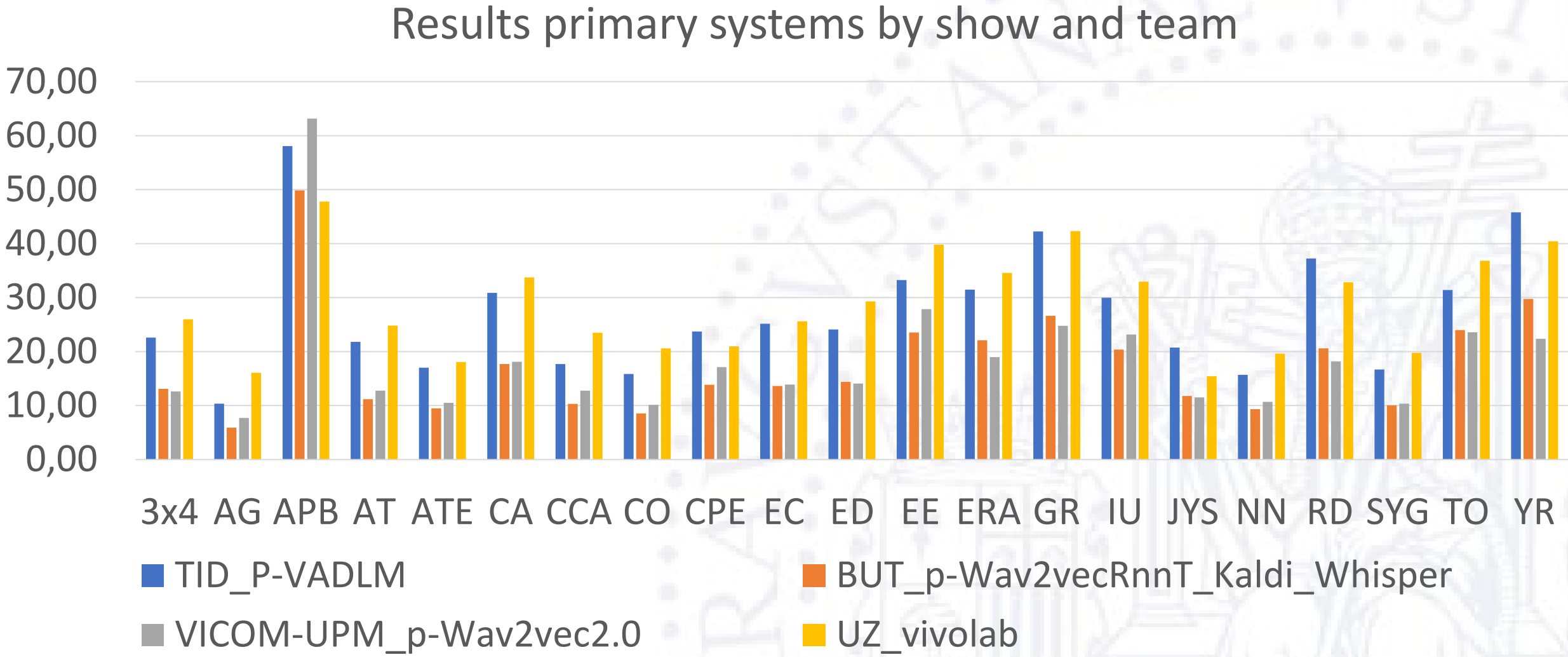
Best 2020 system on 2018 test
11,6%

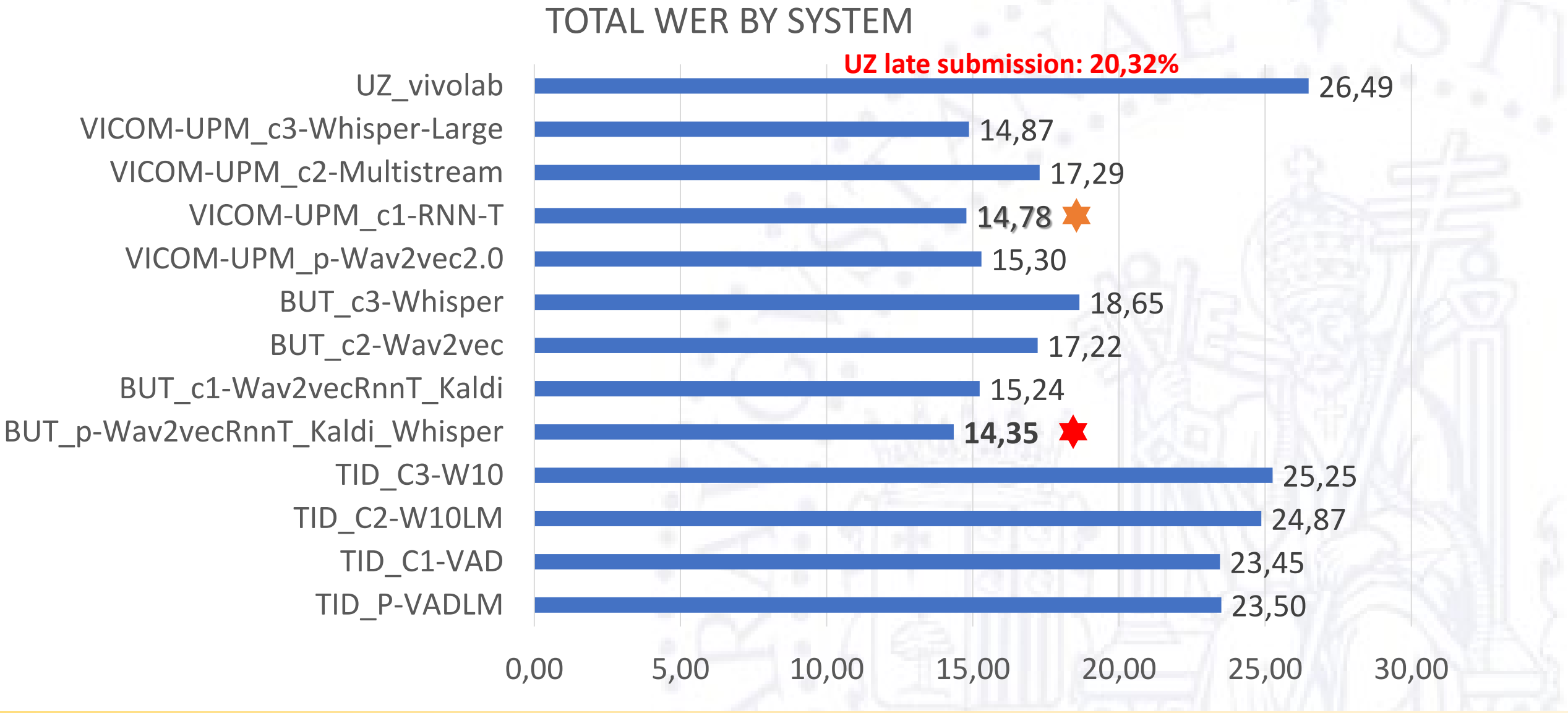


2022 challenge

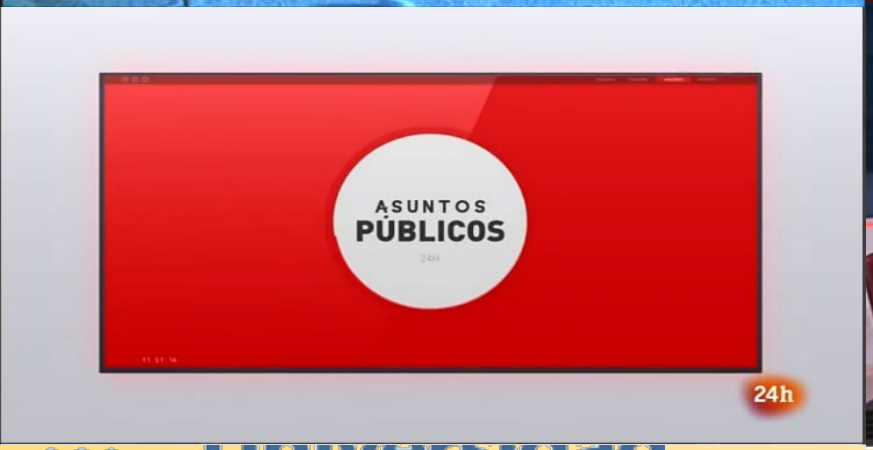


Show	Hours
3x4	2:58:17
A pedir de boca (APB)	3:41:38
Agrosfera (AG)	4:15:20
Aquí la Tierra (AT)	2:46:44
Ateneo (ATE)	1:40:09
Cerámica Popular Española (CPE)	1:02:35
Comando Actualidad (CA)	3:59:29
Conversatorios Casa América (CCA)	1:58:44
Corazón (CO)	3:00:17
El cazador (EC)	3:48:22
Encuestas ruido ambiente (ERA)	2:08:13
Entrevistas en bruto (EE)	3:54:57
España Directo (ED)	4:05:57
Fiction (Grasa-GR, Yrreal-YR, Riders_RD)	3:53:22
Informativos UMATIC (IU)	0:59:49
Jara y Sedal (JYS)	2:29:17
Noticias Nacional (NN)	2:14:32
Saber y Ganar (SYG)	4:28:28
Toros (TO)	0:49:57
21 different shows	**54:16:07**


Results primary systems by show and team
70,00
60,00
50,00
40,00
30,00
20,00
10,00
0,00
3x4 AG APB AT ATE CA CCA CO CPE EC ED EE ERA GR IU JYS NN RD SYG TO YR
TID_P-VADLM
BUT_p-Wav2vecRnnT_Kaldi_Whisper
VICOM-UPM_p-Wav2vec2.0
UZ_vivolab


TOTAL WER BY SYSTEM
UZ late submission: 20,32%
UZ_vivolab
26,49
VICOM-UPM_c3-Whisper-Large
14,87
VICOM-UPM_c2-Multistream
17,29
VICOM-UPM_c1-RNN-T
14,78
VICOM-UPM_p-Wav2vec2.0
15,30
BUT_c3-Whisper
18,65
BUT_c2-Wav2vec
17,22
BUT_c1-Wav2vecRnnT_Kaldi
15,24
BUT_p-Wav2vecRnnT_Kaldi_Whisper
14,35
TID_C3-W10
25,25
TID_C2-W10LM
24,87
TID_C1-VAD
23,45
TID_P-VADLM
23,50
0,00
5,00
10,00
15,00
20,00
25,00
30,00

Speaker Diarization and Identity Assignment Challenge



2022 edition tasks

Evaluate:

- Automatic algorithms for segmenting and clustering speakers in a given audio
- Methods for assigning previously known identities to each speech segment

Evaluation dataset

9 different shows with a total of 25 hours

Show	Hours	
3x4	2:58:17	11,89 %
Agrosfera	4:15:20	17,03 %
Aquí la Tierra	2:46:44	11,12 %
Comando Actualidad	3:59:29	15,97 %
Corazón	3:00:17	12,02 %
España Directo	4:05:57	16,40 %
Ficción (Grasa, Yrreal, Riders)	3:53:22	15,56 %

Diarization Error Rate:

Fraction of speaker time not correctly attributed to a specific speaker

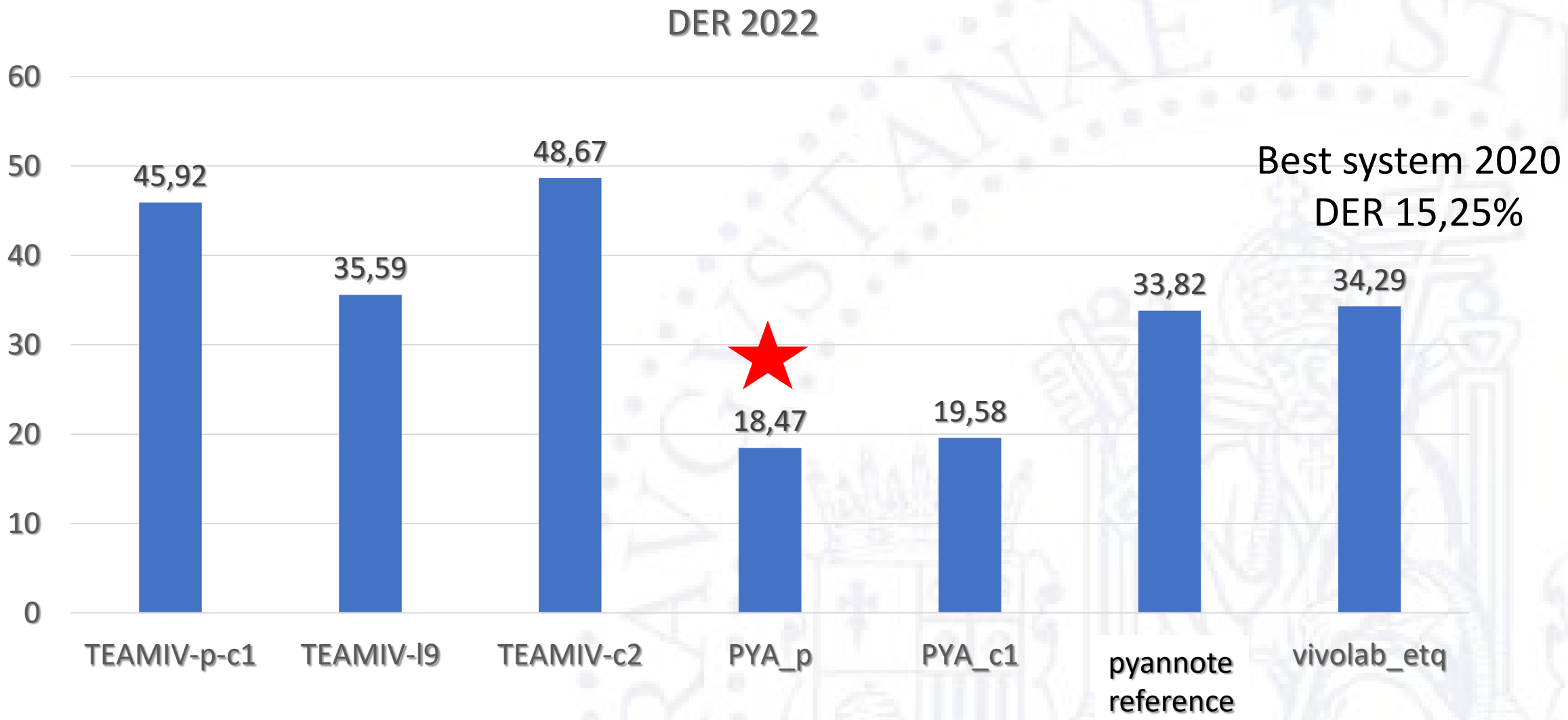
$$DER = \frac{T_{MISS} + T_{FA} + T_{SPK}}{T_{SPEECH}}$$

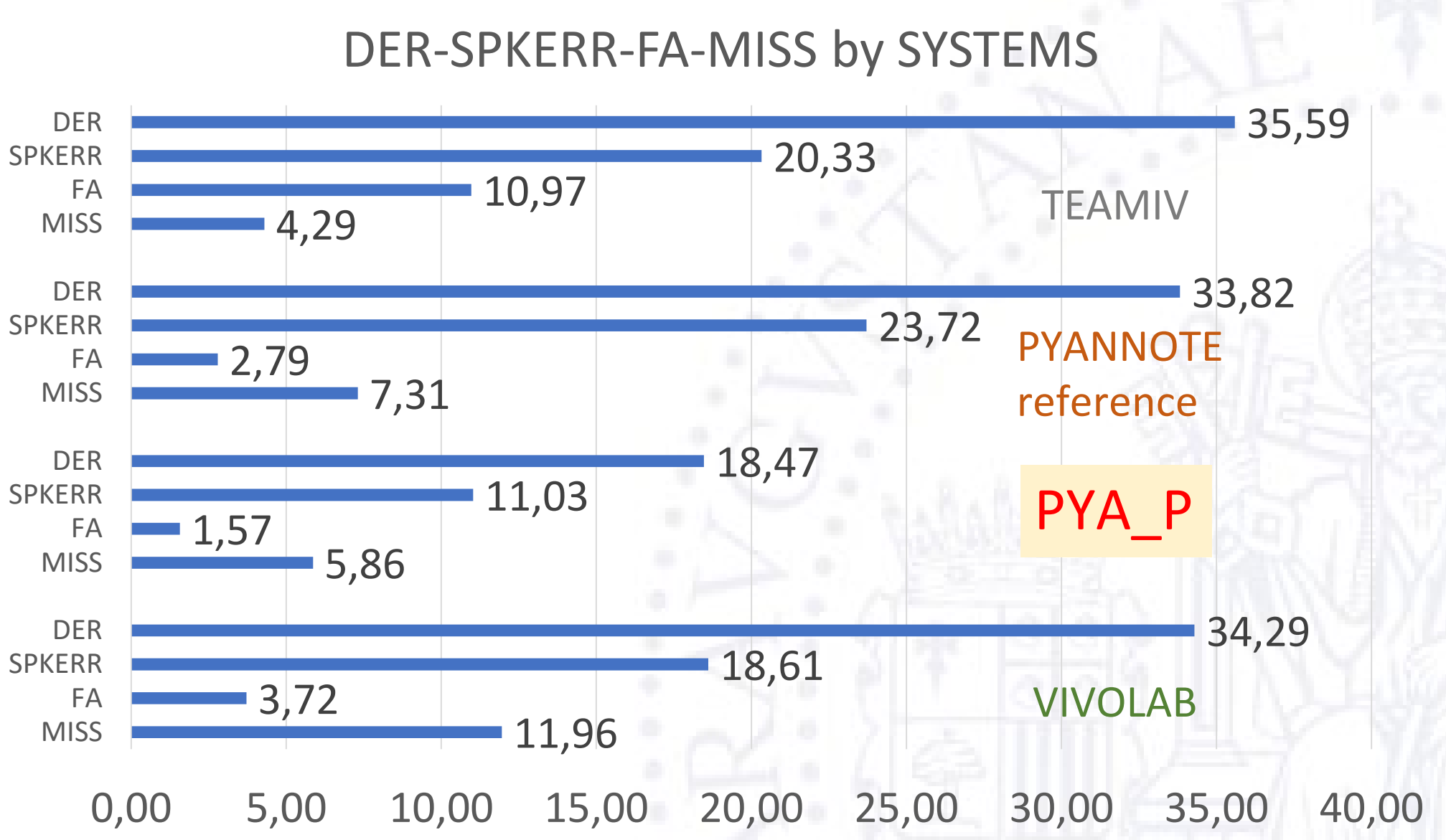
T_{MISS}: Amount of speech considered as non speech

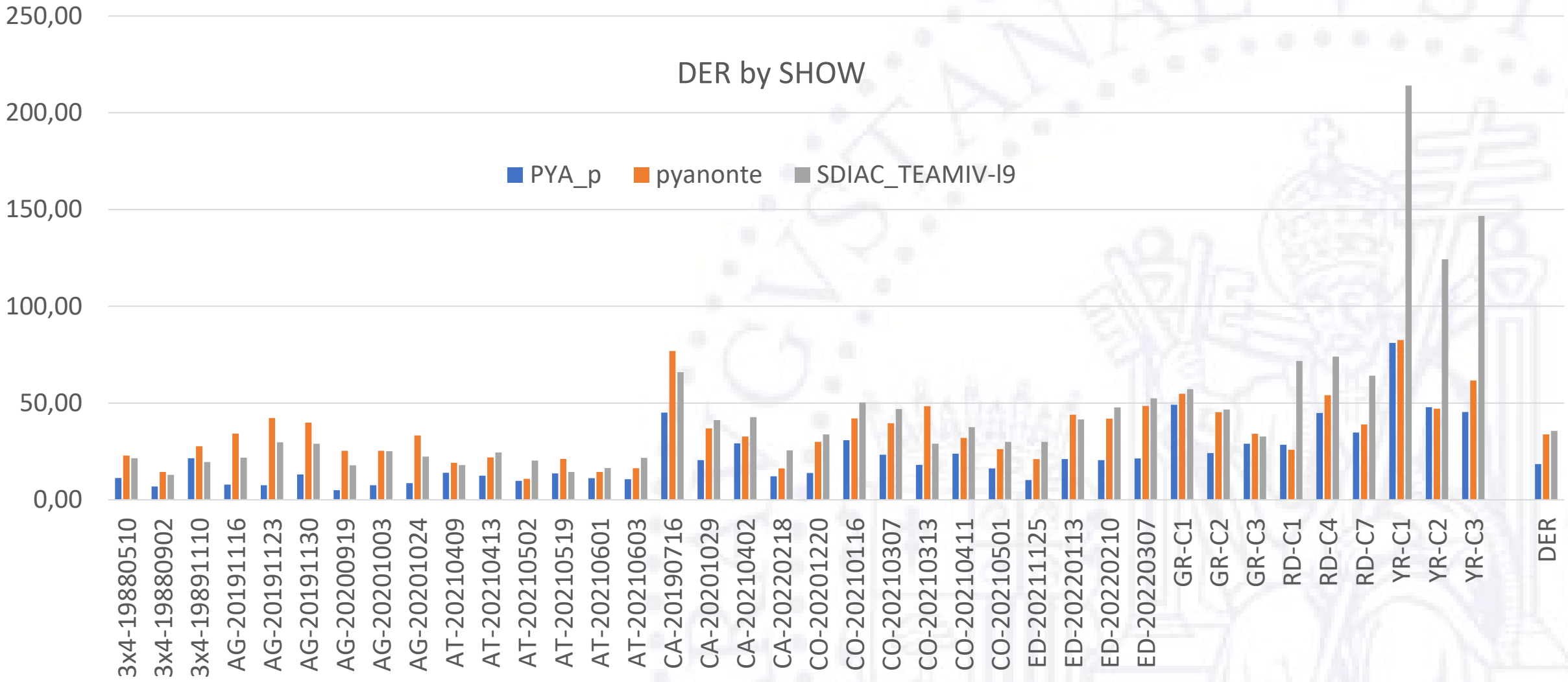
T_{FA}: Amount of non speech considered as speech

T_{SPK}: Amount of speech assigned to a wrong speaker
(contains overlap regions which are evaluated)

250 ms forgiveness collar


DER 2022
60
50
40
30
20
10
0
45,92
35,59
48,67
18,47
19,58
33,82
34,29
TEAMIV-p-c1
TEAMIV-l9
TEAMIV-c2
PYA_p
PYA_c1
pyannote reference
vivolab_etq
Best system 2020
DER 15,25%


DER-SPKERR-FA-MISS by SYSTEMS
DER
SPKERR
FA
MISS
35,59
20,33
10,97
4,29
TEAMIV
33,82
23,72
2,79
7,31
PYANNOTE reference
18,47
11,03
1,57
5,86
PYA_P
34,29
18,61
3,72
11,96
VIVOLAB
0,00
5,00
10,00
15,00
20,00
25,00
30,00
35,00
40,00


DER by SHOW
PYA_p
pyanonte
SDIAC_TEAMIV-l9
250,00
200,00
150,00
100,00
50,00
0,00
3x4-19880510
3x4-19880902
3x4-19891110
AG-20191116
AG-20191123
AG-20191130
AG-20200919
AG-20201003
AG-20201024
AT-20210409
AT-20210413
AT-20210502
AT-20210519
AT-20210601
AT-20210603
CA-20190716
CA-20201029
CA-20210402
CA-20220218
CO-20201220
CO-20210116
CO-20210307
CO-20210313
CO-20210411
CO-20210501
ED-20211125
ED-20220113
ED-20220210
ED-20220307
GR-C1
GR-C2
GR-C3
RD-C1
RD-C4
RD-C7
YR-C1
YR-C2
YR-C3
DER

Assignment Error Rate:

Fraction of speaker time not correctly attributed to a specific speaker

$$AER = \frac{T_{MISS} + T_{FA} + T_{SPK}}{T_{SPEECH}}$$

T_{MISS}: length of speech segments of speaker of interest not attributed to any speaker

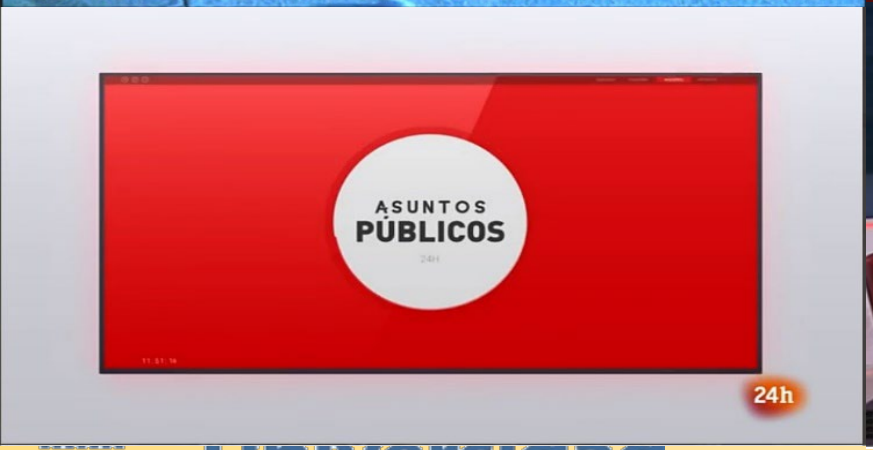
T_{FA}: length of silence segments or speech segments of unknown speakers incorrectly attributed to a certain speaker

T_{SPK}: length of speech segments of speakers of interest attributed to an incorrect speaker

250 ms forgiveness collar

	MISS	FA	SPKERR	AER
TEAMIV_p	1,3	229,7	9,5	240,55
TEAMIV_l3	3,7	160,8	20,9	185,42
TEAMIV_l6	8,3	76,5	5,9	90,62
TEAMIV_l9	**12,4**	**15,3**	**1,2**	**28,88**


Text and Speech Alignment Challenge



(Task 1)TaSAC-ST:

consists of synchronizing the broadcast subtitles created by respeaking of three different TV shows.

22 audios from Agroesfera, Aquí la Tierra and Corazón

A total of 12:10 hours

Metric

Time Error n-th subtitle TE(n)=TEstart(n)+TEend(n)

Program Time Error Metric (PTEM)

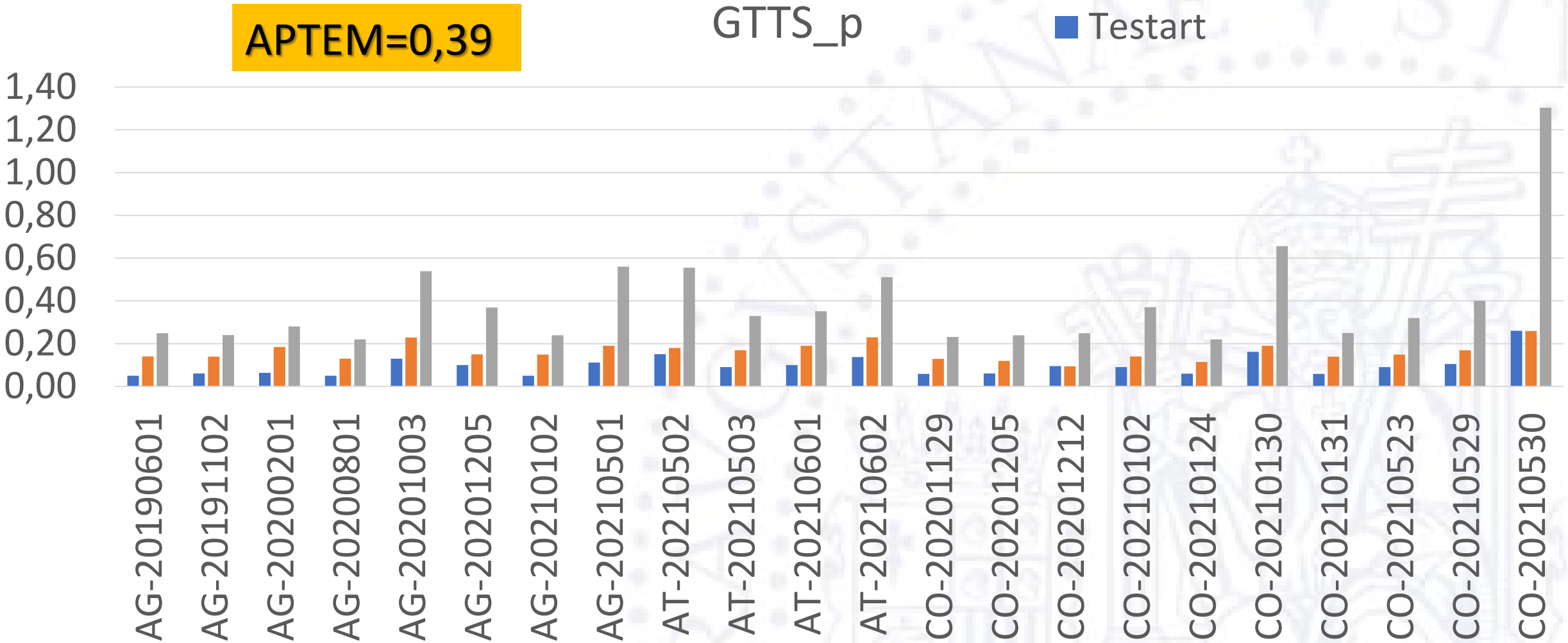
$PTEM(m) = \text{med}\{\text{TE}(1), \ldots, \text{TE}(\text{N})\}$, N: number of subtitles

Average Program Time-Error

$$APTEM = \frac{1}{M}\sum_{m=1}^{M} PTEM(m)$$

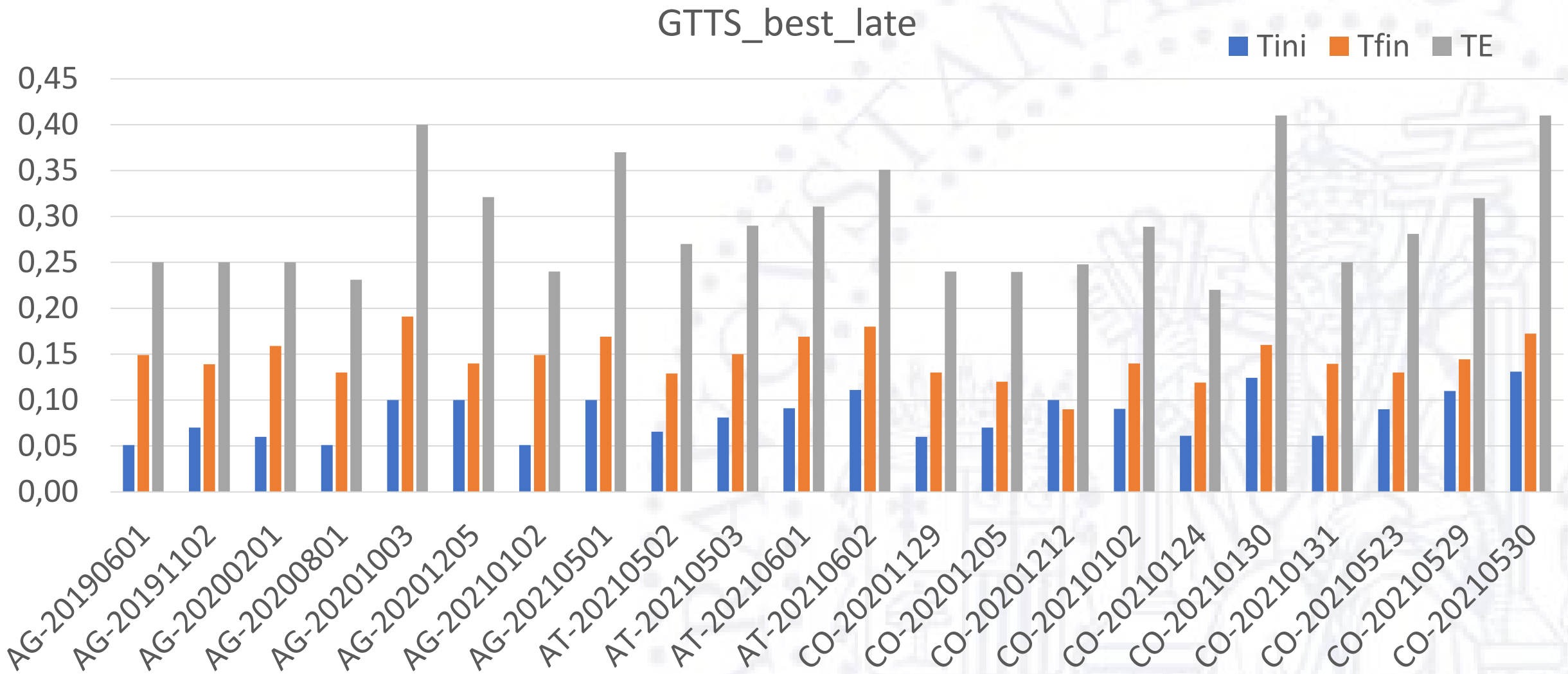
M number of programs

APTEM=0,39


GTTS_p
Testart
1,40
1,20
1,00
0,80
0,60
0,40
0,20
0,00
AG-20190601
AG-20191102
AG-20200201
AG-20200801
AG-20201003
AG-20201205
AG-20210102
AG-20210501
AT-20210502
AT-20210503
AT-20210601
AT-20210602
CO-20201129
CO-20201205
CO-20201212
CO-20210102
CO-20210124
CO-20210130
CO-20210131
CO-20210523
CO-20210529
CO-20210530

APTEM=0,29

APTEstart= 0,08 APTEend= 0,15



- ¿New challenges?
- ¿Codalab competition?
- ¿New schedules?
  - Longer evaluation time
- ¿Give a reward?