The Third DIHARD Diarization Challenge

Neville Ryant[1], Kenneth Church[2], Christopher Cieri[1],
Jun Du[3], Sriram Ganapathy[4], and Mark Liberman[1]

[1]Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA
[2]Baidu Research, Sunnyvale, CA, USA
[3]University of Science and Technology of China, Hefei, China
[4]Electrical Engineering Department, Indian Institute of Science, Bangalore, India

January 23, 2021

Overview

NIST Rich Transcription (RT) evaluation series (2002-2009)

RT series background

- evaluation series run by NIST from 2002 through 2009 that focused (among other things) on speaker diarization
  - → that is, the task of determining "who spoke when" in a recording
- early evaluations focused on diarization of conversational telephone speech and broadcast news
- later evaluations focused on meeting room scenario

Contributions

- substantial performance improvements for diarization of meeting speech
- introduced the diarization error rate (DER) metric, which remains the principal evaluation metric within the field

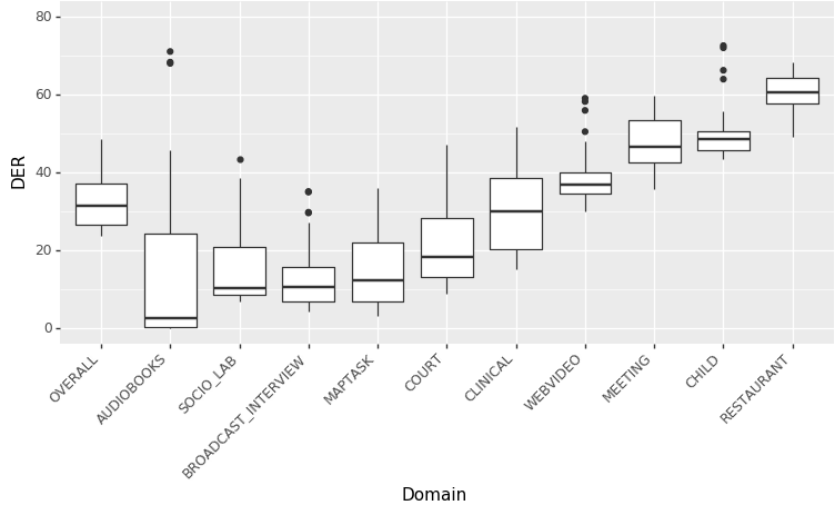
The Wilderness Years (2009-2017)

- 2009-2017: work on diarization continued in many groups
  - ≈48 papers/year included some variant of the term “diarization” in their title
  - ≈409 papers/year mentioned diarization somewhere in their body
- → many clear advances, often in tandem with speaker recognition (e.g., i-vectors, x-vectors, PLDA scoring, improved clustering)
- however, **NO** major evaluations with a diarization component
  - → fragmentation of community
  - different groups focus on different domains/data (e.g., broadcast news, CTS, meeting speech)
  - sometimes inconsistent evaluation procedures
- → difficult to gauge progress

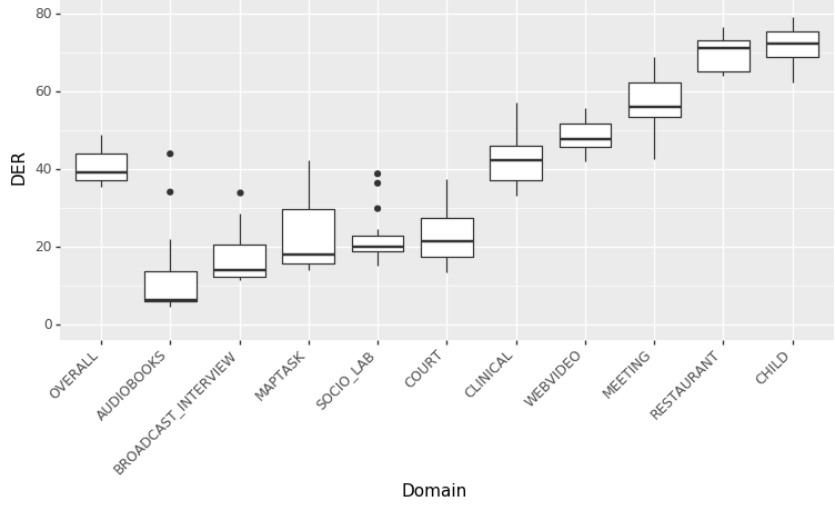
DIHARD I (Interspeech 2018)

- single channel speaker diarization
- 10 diverse domains
- systems evaluated on two tracks:
  - **Track 1:** diarization from reference SAD
  - **Track 2:** diarization from scratch
- primary metric: diarization error rate (DER)
- 13 teams submitted systems
- best DER:
  - Track 1: 23.73% (JHU)
  - Track 2: 35.51% (BUT)

Track 1 – using reference SAD



Track 2 – diarization from scratch

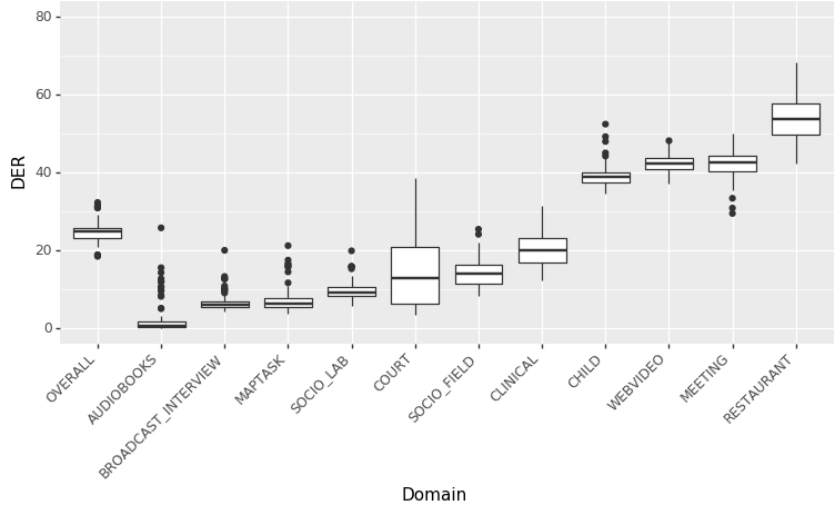


DIHARD II (Interspeech 2019)

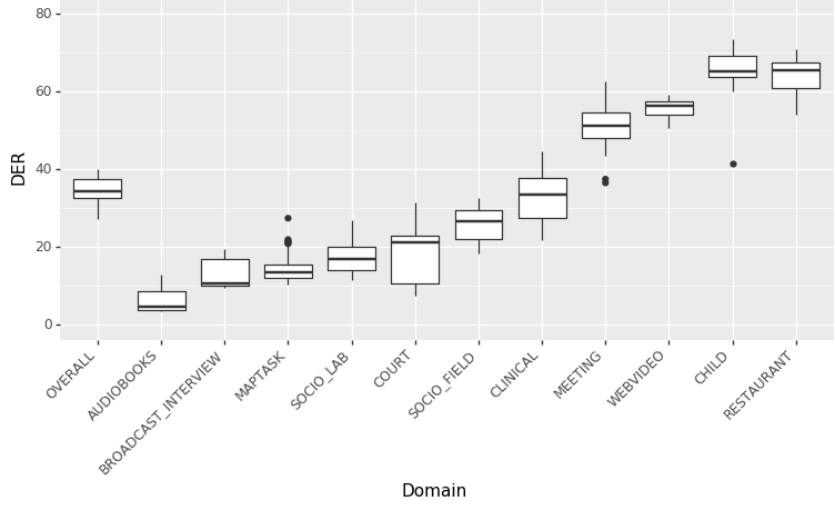
- maintained single-channel tracks from DIHARD I
  - added additional data for some domains
  - vastly improved annotation
- added two tracks evaluating multichannel farfield diarization:
  - CHiME-5 dinner party corpus
- for first time, supplied a baseline system and results
- 21 teams submitted systems

Track	Baseline DER	Best DER (BUT)
Track 1	25.99	18.42
Track 2	40.86	27.11
Track 3	50.85	45.65
Track 4	77.34	58.92

Track 1 – using reference SAD



Track 2 – diarization from scratch



Other recent evaluations

- **CHiME-6** – included diarization as part of Track 2; based on DIHARD II Tracks 3 and 4
  `https://chimechallenge.github.io/chime6/`
- **Fearless Steps I & II** – massively multichannel data from Apollo 11 mission
  `https://fearless-steps.github.io/ChallengePhase2/index.html`
- **VoxSRC-20** – multi-speaker YouTube videos
  `https://www.robots.ox.ac.uk/~vgg/data/voxceleb/competition2020.html`
- **Iberspeech** – diarization of television broadcasts and longitudinal diarization
  `http://catedrartve.unizar.es/albayzin2020.html`

DIHARD III overview

Key changes from DIHARD II

- dropped multichannel diarization; instead, see CHiME-6
- dropped *child language* domain (SEEDLingS)
- additional data for *clinical* domain
- added conversational telephone speech (CTS) domain

NIST collaboration

- for first time, collaborated with NIST via OpenSAT evaluation series
- used NIST platform for all evaluation activities (registration, submission, scoring, etc)
- hope for even closer integration in future challenges

Task

- for each recording:
  - determine how many speakers are present
  - for each speaker, identify all corresponding speech segments
- two tracks:
  - **Track 1** – *Diarization from reference SAD*

    Systems are provided with a reference speech segmentation that is generated by merging speaker turns in the reference diarization
  - **Track 2** – *Diarization from scratch*

    Systems are provided with just the raw audio input for each recording session and are responsible for producing their own speech segmentation
- these tracks are identical to Track 1 and Track 2 from DIHARD I and DIHARD II

Scoring metrics

Diarization error rate (DER)

- sum of three sources of error:
  - missed speech
  - false alarm speech
  - speaker attribution
- computed using:
  - **NO** forgiveness collar
  - explicit scoring of overlapped speech
- primary metric; used for all rankings

Jaccard error rate (JER)

- reference speakers and system speakers optimally paired and the Jaccard index for each pairing computed
- JER is 1 minus the average of these Jaccard indices
- introduced for DIHARD II
- typically, higher than DER, especially when one speaker is dominant

Data

Training data

- no training set is distributed
- participants are free to use any public/proprietary data
- in practice, most teams use some combination of VoxCeleb and CommonVoice

Development/evaluation data

- 5-10 minute duration recordings from 11 conversational domains
- at least 2 hours of audio in each domain
- majority of data segmented manually (and painfully) for DIHARD II
- annotations for new data obtained by forced alignment of turn-level transcriptions

Domains

- *audiobooks*
- *broadcast interview*
- *clinical*
- *courtroom*
- *conversational telephone speech*
- *map task*
- *meeting*
- *restaurant*
- *sociolinguistic field recordings*
- *sociolinguistic lab recordings*
- *web video*

Sources of variation

- interaction type
- recording equipment
- recording environment
- reverberation
- ambient noise
- number of speakers
- speaker demographics
- % speech overlap

Data

Scoring partitions

- **core evaluation set** – a "balanced" evaluation set in which the total duration of each domain is approximately equal
- **full evaluation set** – a larger evaluation set that uses all available selections for each domain; it is a proper superset of the core evaluation set

Set	Part.	# Recordings	Duration (hours)	% speech	% overlap
dev	core	181	23.94	78.43	10.04
	full	254	34.15	79.81	10.70
eval	core	184	22.73	77.35	8.75
	full	259	33.01	79.11	9.35

Core development set

Domain	# speakers	Duration (hours)	% speech	% overlap
AUDIOBOOKS	1.00	2.01	79.30	0.00
BROADCAST INTERVIEW	3.83	2.06	78.65	1.12
CLINICAL	2.00	2.06	60.71	4.66
COURT	6.92	2.08	84.04	1.94
CTS	2.00	2.17	88.83	13.92
MAPTASK	2.00	2.53	67.83	3.01
MEETING	5.36	2.45	93.61	23.24
RESTAURANT	7.17	2.03	87.89	24.03
SOCIOLINGUISTIC (FIELD)	3.50	2.01	72.53	7.83
SOCIOLINGUISTIC (LAB)	2.00	2.67	74.32	5.03
WEB VIDEO	3.97	1.89	74.83	19.91
TOTAL	3.43	23.94	78.43	10.04

Baseline architecture

Overview

- based on LEAP Lab's submission to DIHARD II
- very conventional architecture (i.e., not end-to-end):
  - perform speech activity detection (Track 2 only)
  - divide recording into short overlapping segments
  - extract x-vectors for each segment
  - PLDA scoring
  - agglomerative hierarchical clustering (AHC)
  - VB-HMM resegmentation
- all baseline components as well as recipes for reproducing the official results are distributed via GitHub repo:

  `https://github.com/dihardchallenge/dihard3_baseline`

Acknowledgements

Prachi Singh, Venkat Krishnamohan, and Rajat Varma for all their hard work training and testing!

Baseline architecture

Speech activity detection

- TDNN SAD model based on Kaldi Aspire recipe
  - 5 TDNN layers
  - 2 statistics pooling layers
  - 40-D MFCCs
- DNN outputs smoothed using an HMM with following constraints:
  - min speech duration: 240 ms
  - min non-speech duration: 30 ms
- trained using DIHARD III development data

x-vector extraction

- 512-D x-vectors extracted from overlapping 1.5 sec segments
- x-vectors centered and whitened using DIHARD III development/evaluation set statistics before length normalization
- x-vector extractor trained on VoxCeleb 1+2 using data augmentation

Baseline architecture

PLDA scoring and clustering

- x-vectors scored using Gaussian PLDA model, then clustered using AHC
- prior to PLDA scoring, dimensionality reduced using PCA
- stopping criteria for AHC tuned to minimize DER on the development set
- PLDA model trained using whitened, centered, and length-normalized x-vectors from VoxCeleb 1+2

Resegmentation

- frame-level VB-HMM resegmentation of AHC output
  - 24-D MFCCs (10 ms step)
  - UBM-GMM with 1,024 diagonal components
  - 400-D total variability matrix (V)
- UBM-GMM and V trained using same data as x-vector extractor
- posterior scaling applied to discourage frequent speaker transitions
- run for one iteration

Baseline results

Table: Track 1 evaluation set DER/JER.

Part.	VB-HMM	DER (%)	JER (%)
core	no	21.66	48.10
core	yes	20.65	47.74
full	no	20.75	43.31
full	yes	19.25	42.45

Table: Track 2 evaluation set DER/JER.

Part.	VB-HMM	DER (%)	JER (%)
core	no	29.51	53.82
core	yes	27.34	51.91
full	no	28.00	49.35
full	yes	25.36	46.95

Summary of challenge

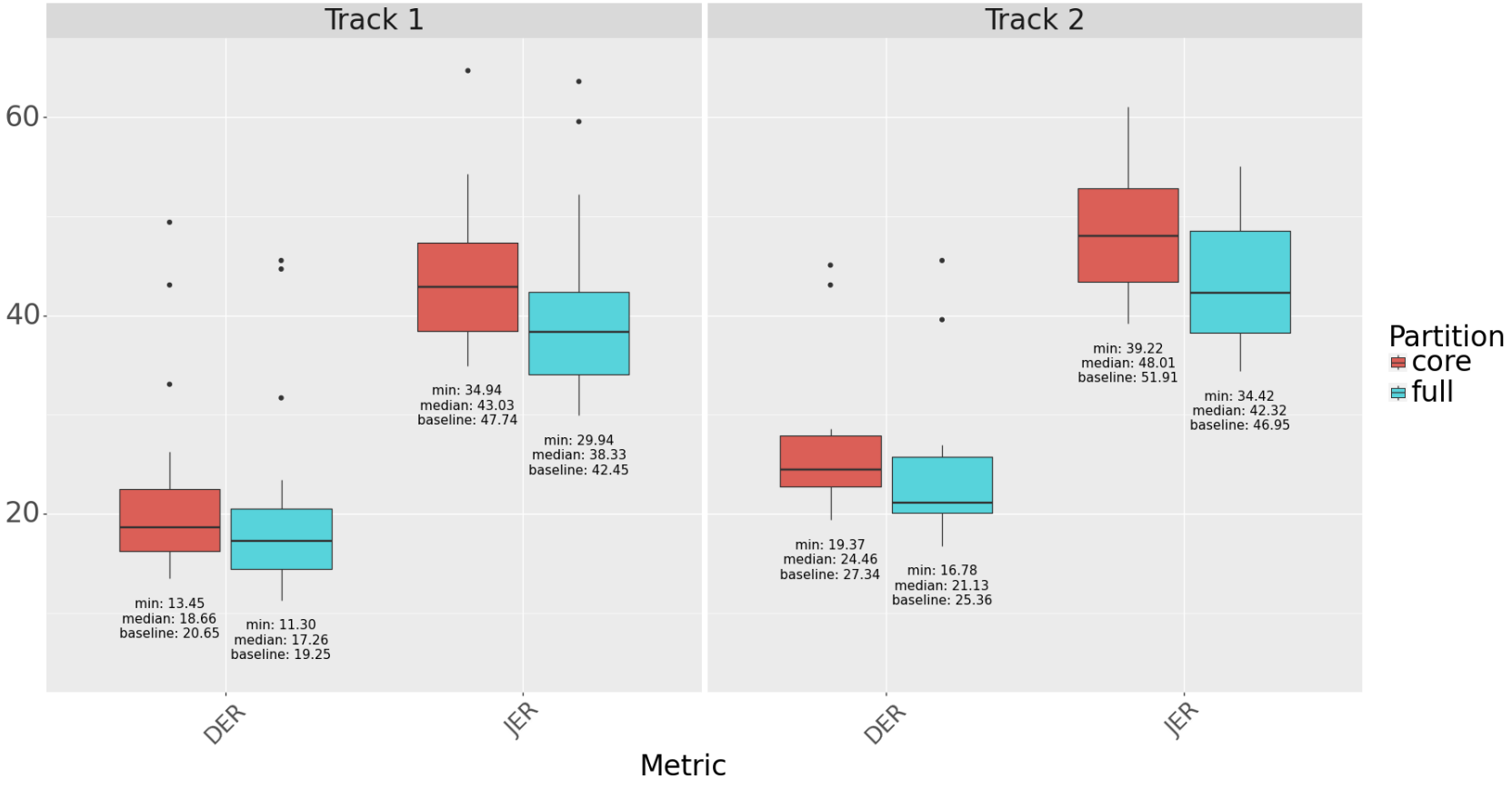
Participation

- 23 teams submitted systems
  - **Track 1:** 23
  - **Track 2:** 14
- participation spanned 10 countries and 3 continents
- actually an improvement over DIHARD II, despite COVID!!!
  - DIHARD I: 13 teams
  - DIHARD II: 21 teams

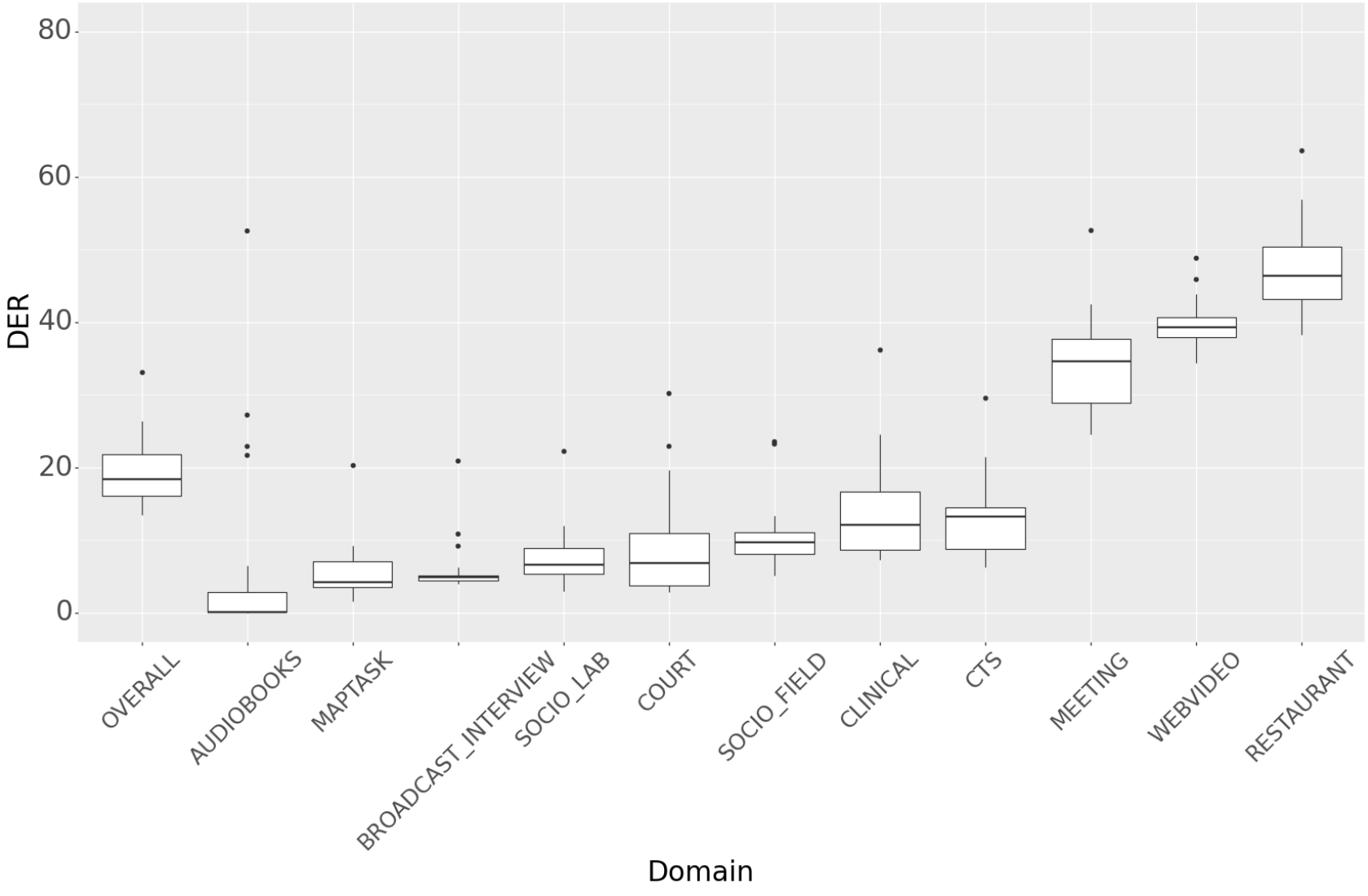
Leaderboards

- Leaderboards available online at:
  `https://sat.nist.gov/dihard3#tab_leaderboard`
- Will be updated in early February with links to system descriptions and full team names.

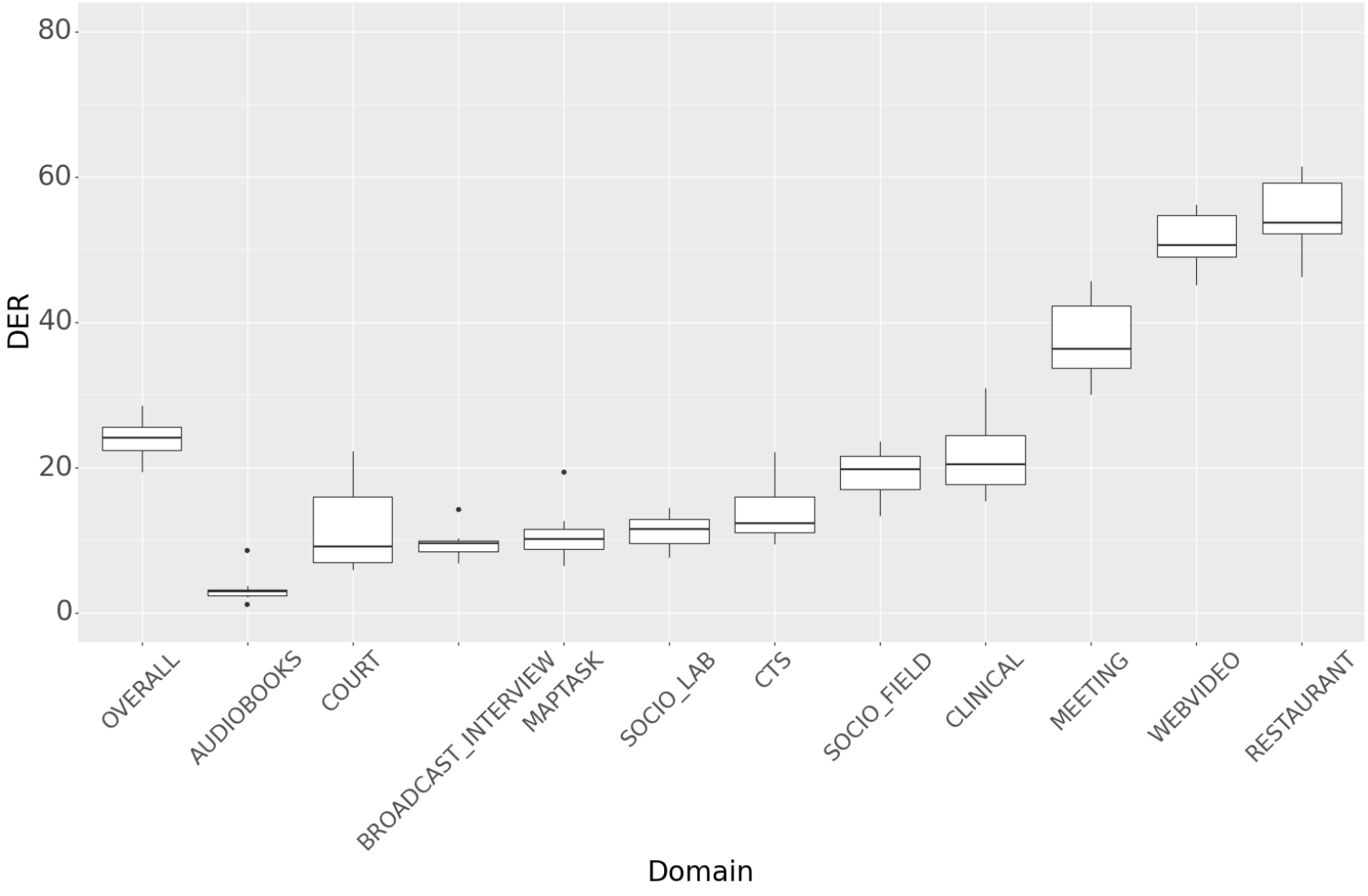
Summary of final leaderboards



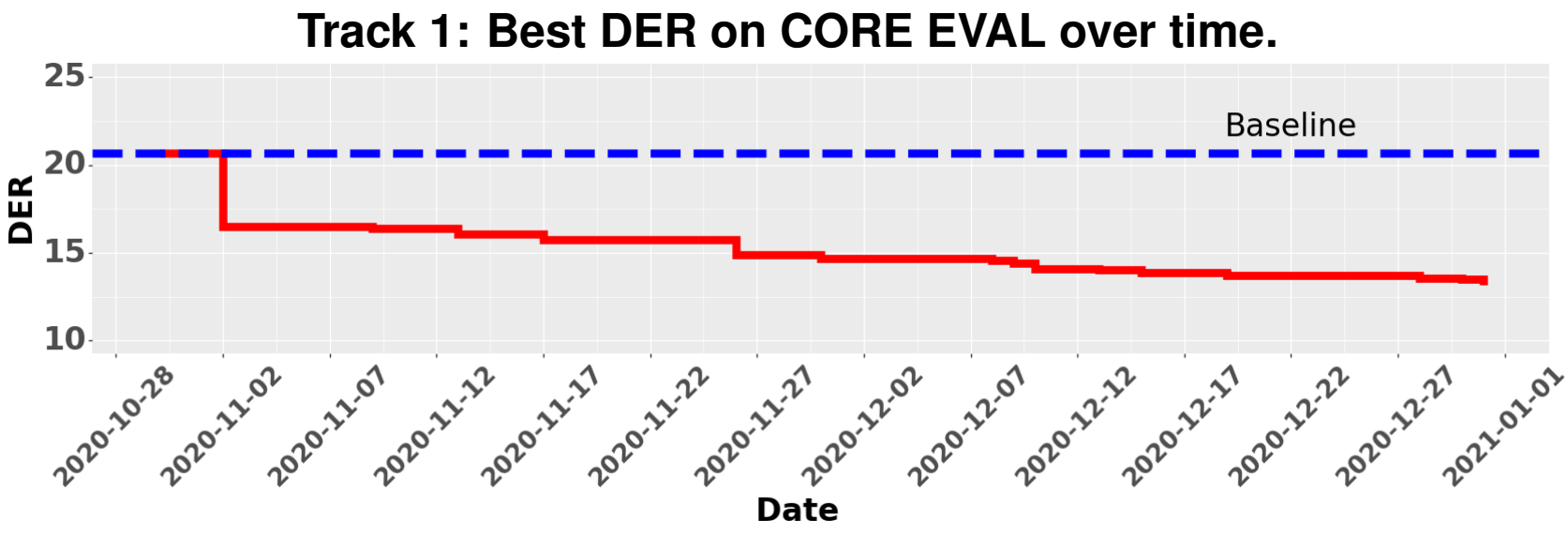
Track 1: Core EVAL set DER by domain

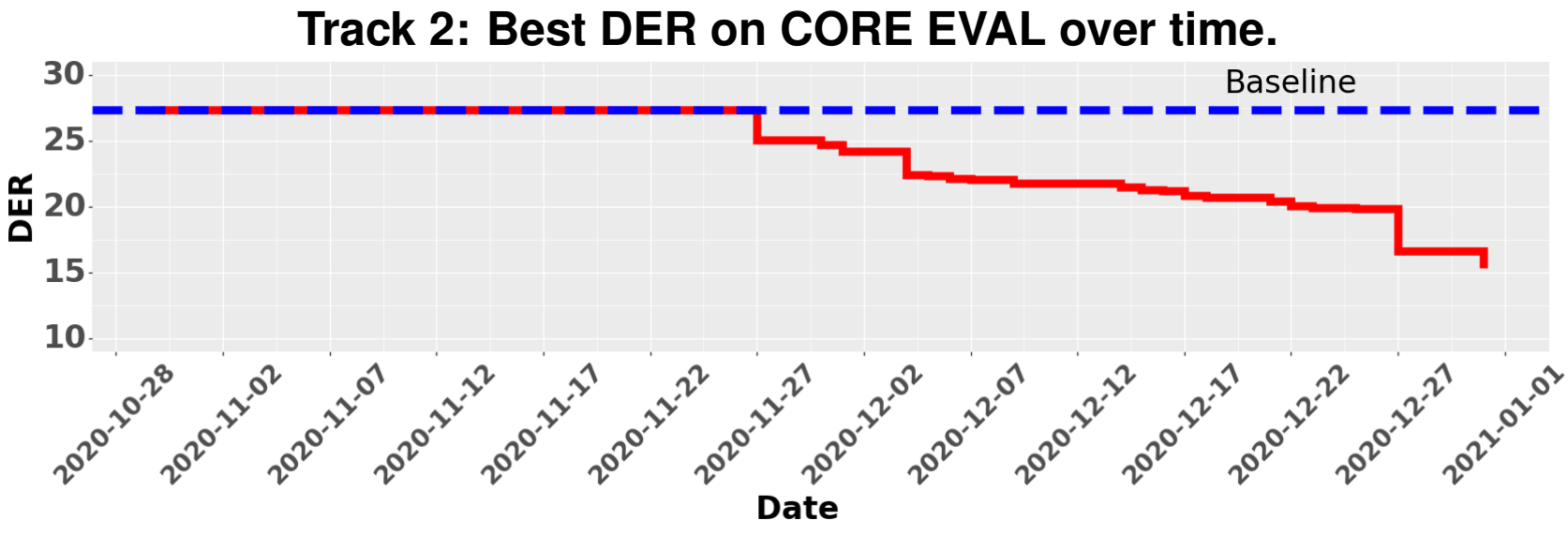


Track 2: Core EVAL set DER by domain



Improvement over course of the challenge





Future plans

Next three months

- release evaluation set answer keys to participants (hopefully in next week)
- publish system descriptions to leaderboards (February)
- post full challenge results to Zenodo (all submissions as well as related metadata) to support work on:
  - system combination methods
  - alternate metrics and evaluation methodologies
- publish full development/evaluation sets via LDC (hopefully by May)
- finish DIHARD II publication; this was derailed by licensing issues

Longer term

- future DIHARDs will occur every two years:
  - yearly schedule too grueling
  - leaves insufficient time for proper post-challenge activities and research
- next DIHARD will be in 2022

Further information

Website

Additional details about the challenge design, baselines, and leaderboards are available via the official website:

`https://dihardchallenge.github.io/dihard3/`

Email

For questions not answered there, or to be added to the mailing list, please send email to:

`dihardchallenge@gmail.com`

6 Backup slides

Diarization error rate

Introduced for RT-03S as the total percentage of reference speaker time that is not correctly attributed to a speaker, where “correctly attributed” is defined in terms of an optimal mapping. Defined as:

$$\text{DER} = \frac{\text{FA} + \text{MISS} + \text{ERROR}}{\text{TOTAL}}$$

where

- *TOTAL* is the total reference speaker time; that is, the sum of the durations of all reference speaker segments
- *FA* is the total system speaker time not attributed to a reference speaker
- *MISS* is the total reference speaker time not attributed to a system speaker
- *ERROR* is the total reference speaker time attributed to the wrong speaker

Jaccard error rate

Speaker specific Jaccard error rate JER_{ref} is computed as:

$$\text{JER}_{ref} = \frac{\text{FA} + \text{MISS}}{\text{TOTAL}}$$

where

- *TOTAL* is the duration of the union of reference and system speaker segments; if the reference speaker was not paired with a system speaker, it is the duration of all reference speaker segments
- *FA* is the total system speaker time not attributed to the reference speaker; if the reference speaker was not paired with a system speaker, it is 0
- *MISS* is the total reference speaker time not attributed to the system speaker; if the reference speaker was not paired with a system speaker, it is equal to *TOTAL*

The Jaccard error rate then is the average of the speaker specific Jaccard error rates:

$$\text{JER} = \frac{1}{N} \sum_{ref} \text{JER}_{ref}$$

Full development set

Domain	# speakers	Duration (hours)	% speech	% overlap
AUDIOBOOKS	1.00	2.01	79.30	0.00
BROADCAST INTERVIEW	3.83	2.06	78.65	1.12
CLINICAL	2.00	4.27	60.04	4.74
COURT	6.92	2.08	84.04	1.94
CTS	2.00	10.17	89.44	13.60
MAPTASK	2.00	2.53	67.83	3.01
MEETING	5.36	2.45	93.61	23.24
RESTAURANT	7.17	2.03	87.89	24.03
SOCIOLINGUISTIC (FIELD)	3.50	2.01	72.53	7.83
SOCIOLINGUISTIC (LAB)	2.00	2.67	74.32	5.03
WEB VIDEO	3.97	1.89	74.83	19.91
TOTAL	3.02	34.15	79.81	10.70

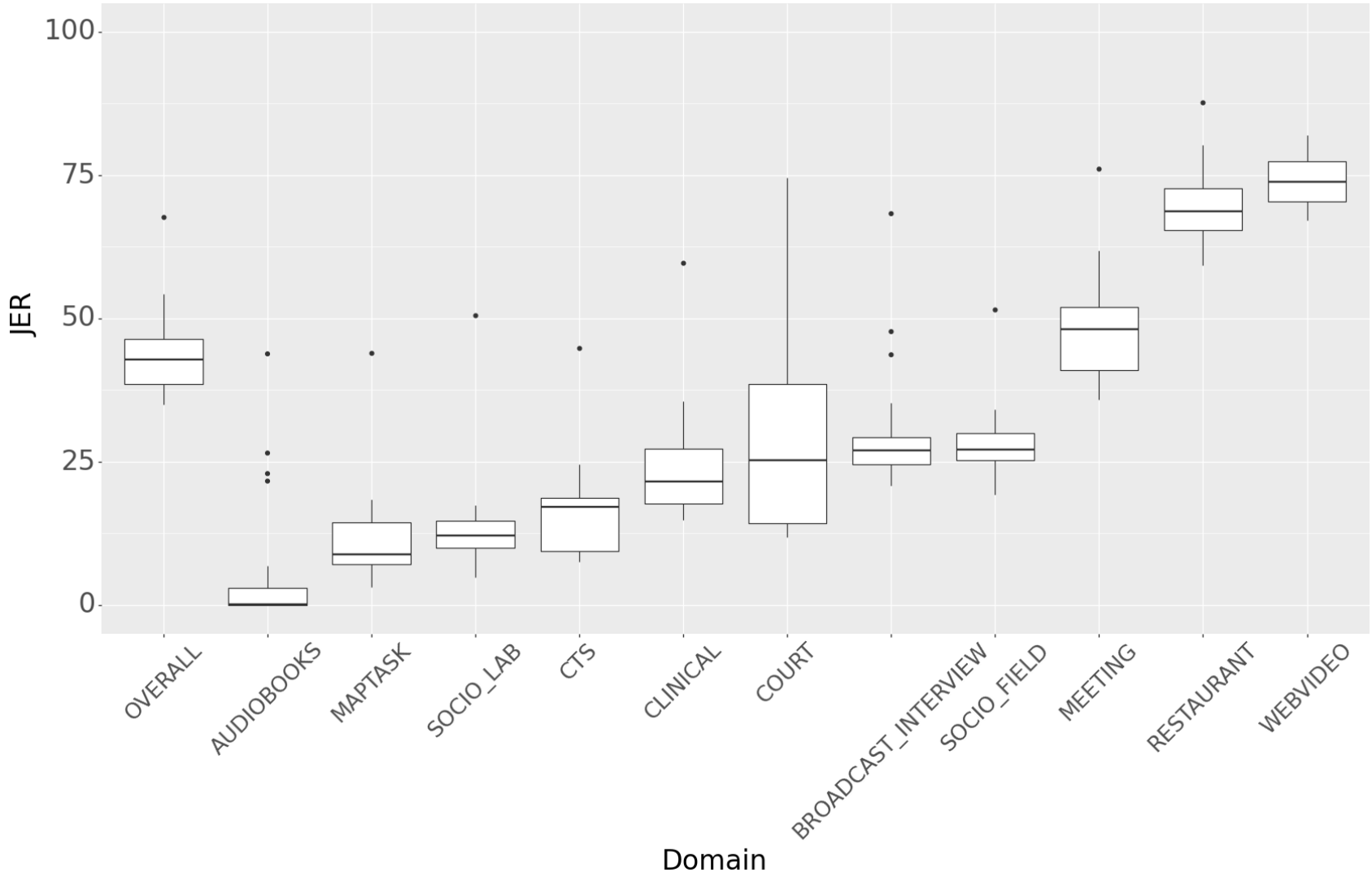
Core evaluation set

Domain	# speakers	Duration (hours)	% speech	% overlap
AUDIOBOOKS	1.00	2.04	77.62	0.00
BROADCAST INTERVIEW	3.67	2.03	77.43	1.61
CLINICAL	2.08	2.08	62.76	3.13
COURT	7.33	2.04	83.13	1.79
CTS	2.00	2.17	88.72	10.92
MAPTASK	2.00	2.07	64.40	1.83
MEETING	3.91	1.87	82.18	21.30
RESTAURANT	6.42	2.06	88.11	26.61
SOCIOLINGUISTIC (FIELD)	2.32	2.27	77.45	4.53
SOCIOLINGUISTIC (LAB)	2.00	2.03	76.00	3.58
WEB VIDEO	4.11	2.07	73.17	17.31
TOTAL	3.24	22.73	77.35	8.75

Full evaluation set

Domain	# speakers	Duration (hours)	% speech	% overlap
AUDIOBOOKS	1.00	2.04	77.62	0.00
BROADCAST INTERVIEW	3.67	2.03	77.43	1.61
CLINICAL	2.04	4.36	61.98	3.22
COURT	7.33	2.04	83.13	1.79
CTS	2.00	10.17	89.08	11.77
MAPTASK	2.00	2.07	64.40	1.83
MEETING	3.91	1.87	82.18	21.30
RESTAURANT	6.42	2.06	88.11	26.61
SOCIOLINGUISTIC (FIELD)	2.32	2.27	77.45	4.53
SOCIOLINGUISTIC (LAB)	2.00	2.03	76.00	3.58
WEB VIDEO	4.11	2.07	73.17	17.31
TOTAL	2.88	33.01	79.11	9.35

Track 1: Core EVAL set JER by domain



Track 2: Core EVAL set JER by domain