The Second DISPLACE Challenge Evaluation Plan

version 1.0

Shareef Babu Kalluri[1], Prachi Singh[1], Shikha Baghel[2], Apoorva Kulkarni[1], Pratik Roy Chowdhuri[2], Deepu Vijayasenan[2], and Sriram Ganapathy[1]

[1]LEAP Lab, Department of Electrical Engineering, Indian Institute of Science, Bangalore, India

[2]Department of Electronics and Communication, National Institute of Technology Karnataka, Surathkal, India

January 2024

1 Introduction

The **DI**riazation of **SP**eaker and **LA**nguage in **C**onversational **E**nvironments (**DISPLACE**) challenge entails a first of kind task to perform speaker and language diarization on the same data containing multi-speaker social conversations in multi-lingual code-mixed speech. In multilingual cultures, social interactions frequently comprise of code-mixed or code-switched speech. Code-mixing occurs when morphemes or words from a secondary language are utilized in a primary language phrase. In contrast, code-switching involves modifying the conversational language itself at the sentence or phrase level. These instances pose significant challenges for speech-based systems, such as speaker and language identification or automatic speech recognition, to extract various analytics for the production of rich transcriptions. The present diarization methods are ill-equipped to handle multilingual conversations in which the same speaker uses numerous languages with different codes.

Inspired by the wide participation in the DISPLACE 2023 challenge and the need for continued research to advance speech technology within natural multilingual conversations, we announced the second season of the DISPLACE challenge. The challenge reflects the theme of Interspeech 2024 “Speech and Beyond- Advancing Speech Recognition and Meeting New Challenges” in its true sense.

The Second DISPLACE challenge aims to

1. highlight new challenges and development in speaker and language diarization for multilingual conversational speech data being evaluated using the underlying dataset and also automatic speech recognition (ASR) in code-mixed/switched multi-accent conversational scenarios.

2. evaluate the performance of submitted systems on this dataset.

Participation in this challenge is open to all who are interested in contributing towards reaching a new milestone in the speaker and / or language diarization and / or ASR areas. For this challenge, we release more than 100 hours of data (both supervised and unsupervised) for development and evaluation purposes. The unsupervised domain matched data is released for participants to use in model adaptation. To the best of our knowledge, no publicly available dataset matches the diverse characteristics observed in the DISPLACE dataset, including code-mixing / switching, natural overlaps, reverberation, and noise. There will be no training data given and the participants are free to use any resource for training the models. A baseline system and an online leaderboard will also be made available to the participants. The results of the challenge will be presented at the Interspeech 2024 conference to be held in Kos Islands, Greece, during 1^{st}-5^{th} September 2024. More information about the challenge can be found on the DISPLACE 2024 website (`https://displace2024.github.io/`).

2 Planned Schedule

Milestones	**Date**
Registration Opens	15 Dec 2023
Development Data Release	10 Jan 2024
Baseline System Release	20 Jan 2024
Leader Board Active	1 Feb 2024
Phase-1 Evaluation Data Release	1 Feb 2024
Registration Closes	10 Feb 2024
Phase-1 Evaluation Closes	28 Feb 2024
System report submission	28 Feb 2024
INTERSPEECH paper submission deadline	2 Mar 2024
INTERSPEECH paper update deadline	9 Mar 2024
Phase-2 Evaluation Data Release	Will be updated soon
Phase-2 Evaluation Closes	Will be updated soon

3 Tasks

This section describes task definition and tracks organized under the DISPLACE 2024 challenge.

3.1 Task Definition

We introduce three tracks in this year's challenge. In Track1 and Track2, the challenge aims to detect and label all speaker and language segments automatically in each conversation, respectively. Track 3 involves automatic speech recognition for multiple Indian languages. Submissions will be evaluated only for speech-based speaker activity regions, including voiced back-channels and fillers, such as yeah, okay, etc. However, non-speech speaker activities, such as laughing, clapping, sneezing, etc., will be excluded from the evaluation. Additionally, small pauses (i.e., $\leq$ 300 ms) taken by a speaker are not considered segmentation breaks and should be a part of the continuous segment. A pause can be described as any segment during which a speaker does not produce any kind of vocalization. Here, vocalization includes speech, speech with errors, vocal sounds (such as laugh, cough, breath, sneeze, lip smacks), non-lexical sounds (i.e., ahh, umm, uh-umm, uh-huh, hmm, huh, ohh, ooo, ahaa, etc.) or any other kind of sound produced by using human sound production system.

3.2 Tracks

The Second DISPLACE challenge organizes three tracks, and the participating teams can register for at least one of the tracks. Each participating team is encouraged to submit their experimental findings and observations to the DISPLACE Challenge at Interspeech 2024 for peer review and subsequent consideration for presentation (and publication) at the conference.

3.2.1 Track-1: Speaker Diarization in multilingual scenarios

This track aims to perform speaker diarization (who spoke when) in multilingual conversational audio data, where the same speaker speaks in multiple code-mixed and/or code-switched languages.

3.2.2 Track-2:Language Diarization in multi-speaker settings

Track 2 aims to perform language diarization (which language was spoken when) in multi-speaker conversational audio data, where the same speaker speaks in multiple languages within the same recording.

3.2.3 Automatic Speech Recognition in multi-accent settings.

Track 3 aims to perform automatic speech recognition in code-mixed/switched multi-accent conversational scenarios.

Each participating team will be provided with a development set (far-field recordings) and a separate baseline system for both Track 1 and Track 2 to enable the design of their own models. For Track 3, we are not releasing the baseline model and participants are free to develop their own model. Subsequently, a blind evaluation set (far-field recordings) will be released. More details about the data can be found in section 5. Each participating team will need to submit their model predictions (in RTTM format for Track 1 and Track 2 and in text format for Track 3) on the blind set to a leader-board interface (setup in Codalab). The leader-board will display the performance of other teams on the same dataset. More details about evaluation rules and protocols can be found in Sections 6 and 7, respectively.

4 Evaluation Score

The systems submitted to both tracks will be evaluated against human reference segmentation. The performance metric for evaluation will be Diarization Error Rate (DER) considering overlapping regions and without tolerance collar for Track 1 and Track 2 and Word Error Rate (WER) for Track 3 (refer section 4). All participants will be required to submit a system description report (Appendix B) to the organizers (refer section 2 for the submission deadline).

4.1 Diarization Error Rate (DER)

DER is the main metric used to evaluate speaker diarization systems in the literature. As described in the NIST Rich Transcription Spring 2003 Evaluation (RT03S) [1], DER is defined as follows:

$$DER = \frac{D_{FA} + D_{miss} + D_{error}}{D_{total}} \tag{1}$$

where,

- D_{FA} represents the total system speaker duration which is not attributed to a reference speaker.
- D_{miss} denotes the total reference speaker duration, which is not attributed to a system speaker.
- D_{error} represents the total system speaker duration attributed to the wrong reference speaker.
- D_{total} denotes the total reference speaker duration, which is represented as the summation of all the reference speakers segments' duration.

The DER metric will be calculated **with overlap** and **without collar**. Here, DER with overlap means that the segments containing the speech of multiple simultaneous speakers are included for evaluation. Also, DER without collar signifies that no tolerance around the actual speaker boundaries is considered during assessment.

4.2 Word Error Rate (WER)

Word Error Rate (WER) is the most common error metric to compute the performance of automatic speech recognition systems. WER is computed by

$$WER = \frac{Insertions + Deletions + Substitutions}{Total\ number\ of\ words\ spoken} \tag{2}$$

where,

[1] `https://web.archive.org/web/20160805233512/http://www.itl.nist.gov/iad/mig/tests/rt/2003-spring/index.html`

- Whenever a word is added that wasn't said in the transcript is treated as **Insertion**.
- Whenever a word is deleted that was said from the transcript is treated as **Deletion**.
- Whenever a word is replaced with a different word to the transcript it is treated as **Substitution**.

4.3 Scoring regions

The scoring region will be the entirety of the recording. For instance, for a 600 seconds long recording, the scoring region will be from 0 second to 600 seconds.

4.4 Scoring tool

The speaker and language diarization system evaluation will be done using dscore (version 1.0.1) script, which is available at:`https://github.com/nryant/dscore`. Each participating team must upload a system output RTTM corresponding to each of the conversations present in the blind evaluation set. The following nomenclature should be followed to score a bunch of system output RTTMs against their corresponding reference RTTMs:

For Track-1 (Speaker diarization in multilingual scenarios)
File naming format for system output RTTMs: $< session\_name > \_SPEAKER\_sys.rttm >$

For Track-2 (Language diarization in multi-speaker settings)
File naming format for system output RTTMs: $< session\_name > \_LANGUAGE\_sys.rttm$

The ASR evaluation will be done using SCTK toolkit available at: `https://github.com/usnistgov/SCTK`. Each participating team must upload a system output transcription text file corresponding to each of the conversations present in the blind evaluation set based on the language information provided.
For Track-3 (Automatic speech recognition in multi-accent settings)
The file naming format for system output is text format: $< session\_name > \_ < Language\_ID > .wrd.trn$

Here, *session_name* should be obtained from the naming convention of audio files, i.e., $< session\_name > .wav$. Refer sub-section 7.1 to understand the result submission requirements.

5 Data

Data description for training, development, and testing set is mentioned in this section.

5.1 Training data

Participating teams are encouraged to use any publicly available and/or proprietary datasets for training and developing the diarization and ASR systems. A detailed description of the data used for training and developing the system must be present in the system description document (see Appendix B) with proper citations to their original resources.

5.2 Development and Evaluation data

The development and evaluation set consist of natural multilingual, multi-speaker conversations. Each conversation (approximately 30 or 60 minutes long) comprises 3-5 participants with proficiency in Indian languages along with English (Indian accent). The data collection paradigm contains a close-talking microphone worn by each speaker and a far-field microphone. The data annotations will be generated using the worn microphone, while the automatic systems will be evaluated on the single-channel far-field audio. The data contains natural code-mixing, code-switching, reverberation, far-field effects, speaker overlaps, short turns, and short pauses. These aspects of the dataset make it unique and challenging. Further, the speech also has a variety of dialects of the same language. Only the development and evaluation data will be released. **Development set can be used for hyper-parameter tuning or training.**

Speakers present in evaluation and development sets are mutually exclusive, i.e., no speaker overlap is present between both sets. The evaluation set contains some languages not present in the development set.

5.3 File format

The audio (far-field) recordings will be released as single-channel wav files sampled at 16 kHz. The naming convention for audio files is " $< session\_name > .wav$". The reference speaker and language segmentations for the development set will be distributed as Rich Transcription Time Marked (RTTM) files for Track 1 and Track 2. A *.segments* and *.trn* files are provided for Track 3. A detailed description of *RTTM*, *.segments* and *.trn* file formats are provided in Appendix A.

6 Evaluation rules

Participation in the DISPLACE challenge is free of cost. Anyone, who follows the evaluation rules mentioned in this plan, can participate in this challenge. The registered team will get access to the development and evaluation data. The DISPLACE challenge is organized as an open evaluation where blind evaluation sets will be sent to each team. The evaluation will be done in two phases, namely, Phase-1 and Phase-2. The evaluation deadlines for these phases are present in section 2. Each team needs to evaluate their systems on the Phase-1, and Phase-2 blind sets locally and upload their output RTTMs to DISPLACE 2024 online leaderboard for evaluation. Refer sub-section 7.1 to know the specified file naming and folder structure for result submission.

As such, the registered team must agree to process the data as per the following rules:

- Each team must submit at least one valid system to the registered track before the end of the evaluation duration of Phase-1. Here, a valid system is described as a submission that contains RTTMs (Track 1 and Track 2) and .trn (Track 3) for all the conversations present in the evaluation set and must pass the validation step during upload.
- The participating teams agree that the evaluation of each test segment should be done based on the information available to the trained system only, i.e., no information obtained from other test segments should be utilized.
- Each team agrees not to probe the evaluation segments through any manual or human means (such as listening, manual segmentation, or transcription generation) before the end of the evaluation period.
- Any automatically derived information, such as domain knowledge, etc., can be used for the development and evaluation set.
- Each team is allowed to make multiple submissions for each of the registered tracks during the evaluation period. For Phase-1, each team can make up to 20 submissions. For each track, the evaluation results will be displayed on a leaderboard for continuous monitoring of the progress. The leaderborad will display the best result obtained by each team. However, teams can access the result of their latest valid submission.

Additionally, the participating teams agree to comply with the following general conditions:

- Each team agrees to submit a system description document (2-4 pages) containing details of the algorithm, data, and computational resources used for the system by the designated deadline (see Section 2). The system description should be submitted in the prescribed format (Appendix B) before the end of the evaluation.
- Each participating team agrees to give authorization for uploading the output RTTMs of their final system (i.e., the one displayed on the leaderboard at the end of the evaluation) on Zenodo. The organizers will upload an archive on Zenodo containing all the system descriptions and system outputs (the final one).

Teams failing to stand by the above rules will not be considered for future evaluations. Their registrations will not be accepted until they are committed to participating fully.

7 Evaluation protocol

The system evaluation will be performed through a leaderboard, hosted and maintained at the CODALAB web interface, for active progress monitoring.

7.1 Setting up an evaluation account

Each participating team must create a CODALAB account for evaluating their system output. The account can be created using the following link:

`https://codalab.lisn.upsaclay.fr/accounts/signup/`

While signing up for an account, each team must use their **team name as username and contact email address as provided in the registration form**. By signing up for the evaluation account, each team agrees to the evaluation terms and conditions.

7.2 Evaluation registration

After successfully creating a CODALAB account, each team should participate in the DISPLACE 2024 challenge evaluation using the following link:

`to_be_updated_soon`

In *Participate* section, check the box for "I accept the terms and conditions of the competition" and click on *Register*.

7.3 Results submission

After successful registration, each team can start evaluating its system. For both tracks, results submission should be made by following the rules mentioned here. The submitted results will be first passed through a validation step. The evaluation will be done only if the validation step is successfully passed.

7.3.1 Preparing the submission archive

The validation step ensures that the following rules are satisfied.

1. Each submission should be made through a compressed file. **Only ".zip" or ".tgz" compressed file formats are allowed.**

2. The submitted compressed file must have the following directory structure:

   For Track-1 (Speaker diarization in multilingual scenarios)
   SPEAKER.zip/ < session_1_name > _SPEAKER_sys.rttm >
   SPEAKER.zip/ < session_2_name > _SPEAKER_sys.rttm >
   .
   .
   SPEAKER.zip/ < session_N_name > _SPEAKER_sys.rttm >

   For Track-2 (Language diarization in multi-speaker settings)
   LANGUAGE.zip/ < session_1_name > _LANGUAGE_sys.rttm >
   LANGUAGE.zip/ < session_2_name > _LANGUAGE_sys.rttm >

.

.

$LANGUAGE.zip/ < session\_N\_name > \_SLANGUAGE\_sys.rttm >$

For Track-3 (Automatic Speech Recognition in multi-accent settings)
$LANGUAGE\_ID.zip/ < session\_name > \_LANGUAGE\_ID.wrd.trn >$
$LANGUAGE\_ID.zip/ < session\_name > \_LANGUAGE\_ID.wrd.trn >$

.

.

$LANGUAGE\_ID.zip/ < session\_name > \_LANGUAGE\_ID.wrd.trn >$

NOTE: The compressed file must **not** contain the following structure:
$SPEAKER.zip/ < Directory > / < session\_1\_name > \_SPEAKER\_sys.rttm >$
$LANGUAGE\_ID.zip/ < Directory > / < session\_name > \_LANGUAGE\_ID.wrd.trn >$

3. The submission file must contain a single output RTTM corresponding to each of the conversations present in the evaluation data.

If any of these rules are not satisfied, the submission will not be considered valid, and evaluation will not be performed.

8 Updates

Any changes to this evaluation plan will be notified through the mailing list and the challenge website (`https://displace2024.github.io/`).

Appendix A:

RTTM File Format Specification

The output of the systems should be given as Rich Transcription Time Marked (RTTM) files. An RTTM file is a text file with one speaker/language turn per line. Every line contains the following ten space-delimited fields:

1. **Type** – segment type;
   It should be "SPEAKER" for speaker turns and "LANGUAGE" for language turns.
2. **File ID** – file name;
   It represents the file name (e.g., B014 for all the speaker/language segments present in B014.wav).
3. **Channel ID** – channel (1-indexed) that turn is on; should always be 1
4. **Turn Onset** – onset of turn in seconds from beginning of recording
5. **Turn Duration** – duration of speaker/language turn (in seconds)
6. **Orthography Field** – should always by <NA>
7. **Speaker Type** – should always be <NA>
8. **Speaker ID / Language ID** – Relative speaker/language ID of the turn;
   It should be unique within the scope of each file. (Ex: "S1"− > for speaker ID and "L1" − > for language ID)
9. **Confidence Score** – system confidence;
   It denotes the probability that the information is correct. It should always be <NA>.
10. **Signal Lookahead Time** – should always be <NA>

For example:

1- For speaker diarization

SPEAKER B026 1 2.262 39.073 <NA> <NA> S1 <NA> <NA>
SPEAKER B026 1 25.363 1.364 <NA> <NA> S2 <NA> <NA>
SPEAKER B026 1 32.962 0.449 <NA> <NA> S3 <NA> <NA>

2- For language diarization

LANGUAGE B026 1 2.262 2.094 <NA> <NA> L1 <NA> <NA>
LANGUAGE B026 1 4.356 0.945 <NA> <NA> L2 <NA> <NA>
LANGUAGE B026 1 5.301 0.358 <NA> <NA> L1 <NA> <NA>

LANGUAGE B026 1 5.659 1.228 <NA> <NA> L2 <NA> <NA>

Transcription File Format Specification (.trn)

The transcription file contain segment-wise ground truth transcripts in the following format:

$Transcription\ (< SEGMENT\_ID1 >)$
$Transcription\ (< SEGMENT\_ID2 >)$
.
.
.
$Transcription\ (< SEGMENT\_IDn >)$

For example:
Is it a part of your tradition? (juot-juot_en_214880_216420)
I can see all of them. (juot-juot_en_217988_219974)
So you might have seen Diwali. (juot-juot_en_262190_264218)

Note: It is must to keep the same segment id for the corresponding system transcripts.

Appendix B: System descriptions

Each participating team with at least one valid submission must submit a system description document. This document should contain sufficient details about the algorithm, data, and computational resources so that the results can be reproducible. To maintain a consistent presentation and format, participants must use the Interspeech-2024 conference proceedings template, which can be found at `https://interspeech2024.org/author-resources/`.

The system description document must contain the following structure:

- Section 1: Authors
- Section 2: Abstract
- Section 3: Notable highlights
- Section 4: Data Resources
- Section 5: Detailed description of the algorithm
- Section 6: Results on the development set
- Section 7: Hardware requirements

Section 1: Authors
List of researchers whose contribution you wish to acknowledge.

Section 2: Abstract
A short description of the submission.

Section 2: Notable highlights
A brief description of any kind of novelty or important observation made during experimentation. For example, differences among submitted systems, novel approaches, or features that contributed significantly to system performance improvement.

Section 5: Detailed description of the algorithm
This section must provide sufficient details of each and every component of the system so that a fellow researcher can understand, re-implement the whole system, and reproduce the same results. No need to provide a detailed description of standard components, such as CNN, LSTM, etc. Hyperparameter tuning (if performed) must be thoroughly explained and accompanied by the final obtained hyperparameters. Each of the main sub-tasks of the system should be explained under separate sub-sections. For instances,

- Signal processing – e.g., signal enhancement, denoising, source separation
- Acoustic features – e.g., MFCCs, PLPs, mel fiterbank, PNCCs, RASTA, etc.

- Speech activity detection – relevant for both the tracks
- Segment level representation – e.g., i-vectors, d-vectors
- Speaker estimation – description of how the number of speakers was estimated (if performed)
- Clustering method – e.g., k-means, agglomerative
- Resegmentation details

Section 6: Results on the development set
Each team must report the performance of their submitted systems on the entire development set. The results should be reported in terms of DER using the official scoring tool (available at:`https://github.com/nryant/dscore`). Teams are also encouraged to report the performance of other experiments, such as results on additional datasets or systems. The performance of main system components that are believed to result in significant performance improvements should be clearly quantified.

Section 7: Hardware requirements
Hardware required during the system training and testing should be explicitly mentioned. Here are some examples of the hardware requirement:

- Total number of CPU cores used
- Description of CPUs used (model, speed, number of cores)
- Total number of GPUs used
- Description of GPUs used (model, single precision TFLOPS, memory)
- Total number of TPUs used
- Generations of TPUs used (e.g., v2 vs v3)
- Total available RAM
- Used disk storage
- Machine learning frameworks used (e.g., PyTorch, Tensorflow, CNTK)

Teams must report the system execution time required to operate the whole development set.