PS-06 – Language Agnostic Speaker Identification & Diarization; and subsequent Transcription & Translation System

1. The Grand Challenge (GC) Problem Statement (PS) 06 aims to develop an integrated audio processing system that transforms spoken audio into structured, multilingual and textual insights. Given an input audio file, the system will perform the following tasks in offline mode (i.e. without any dependency on Internet): -

(a) **Speaker Identification:** Match each speaker segment to a known identity when enrolment data is available.
(b) **Speaker Diarization:** Segment the audio by identifying boundary between different speakers.
(c) **Language Identification:** Detect the language spoken in each segment, supporting multilingual and code-switched audio.
(d) **Automatic Speech Recognition:** Convert each speaker's speech into accurate text in spoken script.
(e) **Neural Machine Translation:** Translate the transcribed text into English language, preserving speaker level segmentation.

2. **The solution should be able to handle audio files with:-**
(a) Sample rate: 8k to 48 k
(b) Bit depth : 8 to 32 bits
(c) File types : should cover at least wav, mp3, ogg, flac.
(d) SNR : 5 db or better

3. **Languages to be covered for transcription and translation:-**
(a) Stage 1: English, Hindi and Punjabi
(b) Stage 2: English, Hindi, Punjabi, Bengali, Nepali, and Dogri.
(c) Stage 3: English, Hindi, Punjabi, Bengali, Nepali, Dogri, and five more languages, which will be disclosed to stage 3 participants before the start of stage 3.

4. Participant are advised to use their own or open-source (e.g. AI Kosh) data sets to train their model. However, while training their model, they should condition the data sets (audio files) as per para 2 above as well as the data set should include natural overlaps, code-switching, reverberations and noise.

5. Mock data set (audio files with ground truth) will be released on 15 Sep 25 for self-assessment by the participant.

6. Different data sets for evaluation (Shortlisting data sets) will be released three times on 20th, 25th and 31st Oct 25 at 1100 hrs and results has to be submitted by the participants within 06 hrs from release of these data sets. The shortlisting will be made based on the best of these three results on all five sub-problem statements

(SID, SD, LID, ASR & NMT). All the participants are required to submit their results in the format mentioned at **Appendix 'A'.** The list of shortlisted participants will be published along with the cutoff score as per the evaluation criteria mentioned below. Participants individual scores will be shared over the email.

Sr. No.	Criteria	Weightage (%)	Remarks
(a)	Speaker Identification (Top-1 Identification Accuracy)	15%	• Refer **Appendix 'A'** for result submission and **Appendix 'B'** for Evaluation formulas.
(b)	Speaker Diarization (Diarization Error Rate)	20%	
(c)	Language identification (Diarization Error Rate).	20%	
(d)	Automatic Speech Recognition (Word Error Rate)	30%	
(e)	Neural Machine Translation (BLEU score)	15%	

7. Based on the shortlisting results, at most 15-20 shortlisted solutions will be called in for stage 1 final evaluation **at IIT Delhi** in Nov 25. The number of shortlisted participants may vary based on the overall performance at the discretion of the Jury for this Problem Statement.

8. For stage 1 final evaluation, the shortlisted participants are required to setup their solution at our infrastructure at IIT Delhi. The solution should be able to work in offline mode (i.e. without any dependency on internet connection). The holdout data set will be provided and results will be generated in the presence of the jury. These results will be used for finalizing stage-1 winners as per the following evaluation criteria :-

Sr. No.	Criteria	Weightage (%)	Remarks
(a)	Speaker Identification (Top-1 Identification Accuracy)	10%	Refer **Appendix 'A'** for result submission and **Appendix 'B'** for Evaluation formulas.
(b)	Speaker Diarization (Diarization Error Rate)	15%	
(c)	Language identification (Diarization Error Rate).	15%	
(d)	Automatic Speech Recognition (Word Error Rate)	20%	
(e)	Neural Machine Translation (BLEU score)	10%	

(f)	Proposed architecture & Approach adopted	20%	Based on Presentation & Interview of shortlisted Participant.
(g)	Technical capabilities of the team, Ability of the team to complete overall objectives	10%	

Note: The submitted solutions/models will be evaluated only for speech-based speaker activity regions, including voiced back-channels data 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. ≤ 500 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.

9. All the participants are requested to pack their complete solution in docker or virtual environment format (including all dependencies) in a single compressed file and submit only MD5 hash of the same along-with their shortlisting data set results mentioned at para 6 above. Only the APIs are to be exposed for evaluation by GC team. Please note that the participant may be asked to use this compressed file (docker\virtual file) along with MD5 hash to verify their shortlisting data set results at any stage of Grand Challenge. Participant are not required to submit their solution with shortlisting data set results. Hence, they are advised to carefully & securely save this compressed file. Participant's inability to produce compressed file with matching MD5 hash will result in disqualification.

10. The compute\memory resources used for deploying the model for evaluating the shortlisting data set provided by GC Team should be disclosed to the GC team clearly indicating the servers (processor types, cores, Memory) and GPU (type and size). Kindly note that these details will be used to verify the system performance criteria during offline evaluation.

11. There should not be any inconsistency/ disparity in the results during the evaluation or shortlisting stages. Any unfair practices will lead to disqualification and barring the participants apart from other actions.

12. The evaluation criteria for stage 2 & stage 3 would be similar to that for stage 1 and would be released to qualified participants before start of the stages.

13. **Updates:** Updates will be notified on the GC website.

14. **Sessions with Mentors\Experts :**

(a) For Stage-1, the organisers plan to meet participants via online meet or email to resolve their doubts, if any. This provision will be made active from 15th Aug 2025 and details regarding interaction will be shared on this website. Kindly keep viewing this website regularly for updates on this.

(b) There will be sessions with Mentors\Experts in Stage-2 and Stage-3 for the willing selected participants to help them in achieving the best solutions.

Appendix 'A'

Submission of Results

1. The file formats for sharing results is as given below:

   (a) Speaker identification: CSV format with segmented time stamps for identified speaker.

   (b) Speaker diarization: CSV format with segmented time stamps for each speaker.

   (c) Language identification: CSV format with segmented time stamps for each language.

   (d) Automatic Speech Recognition (ASR): TRN file with segmented time stamps, transcript in accordance with language diarization.

   (e) Neural Machine Translation (NMT): TXT file with segmented time stamps, translation in accordance with ASR file.

2. For evaluation, the evaluation data set (audio files) will be provided in a compressed file. Evaluation data set will have a ReadMe file with brief description of data, Audio Files and other relevant files with following naming convention:

   Compressed file Name : ps6_<evalutaion_id>.<ext>

   Audio File Name: ps6_<evaluation_id>_<audio_file_id>.<ext>

      Evaluation ID: To be issued by GC team in succession of release of audio data set.

      Audio File ID : Unique ID to be issued by GC team for each audio file in evaluation data set.

   Example:

   (a) For first evaluation data : ps6_01_001.wav, ps6_01_002.mp3 etc

   (b) For second evaluation data : ps6_02_001.mp3, ps6_02_002.ogg etc

3. The result files are required to be submitted in following convention for each sub-PS :

   (a) Speaker identification Output: SID_< evaluation_id >.csv

   (b) Speaker Diarization Result: SD_<evaluation_id>.csv

   (c) Language identification/ diarization Output: LID_< evaluation_id >.csv

   (d) ASR Output: ASR_< evaluation_id >.trn

(e) NMT Output: NMT_< evaluation_id >.txt

4. Each of the above files must contain result data arranged in following format:

(a) Note: TS is time stamp in seconds from start of audio file.

(b) Speaker identification Output: SID_< evaluation_id>.csv

Audio File Name, speaker ID, confidence score (in %), start TS, end TS

Example:

ps6_01_001.wav, ID1, 95, 100.006, 120.002

ps6_01_001.wav, ID2, 93, 118.080, 200.256

ps6_01_002.wav, ID2, 90, 234.130, 339.786

ps6_01_003.wav, ID1, 85, 014.540, 058.954

(c) Speaker Diarization Output: SD_<evaluation_id>.csv

Audio File Name, speaker, confidence score (in %), start TS, end TS

Example:

ps6_01_001.wav, speaker1, 95, 100.006, 120.002

ps6_01_001.wav, speaker2, 93, 118.080, 200.256

ps6_01_001.wav, speaker1, 90, 234.130, 339.786

ps6_01_002.wav, speaker1, 90, 234.130, 339.786

ps6_01_003.wav, speaker1, 85, 014.540, 058.954

ps6_01_003.wav, speaker2, 75, 060.240, 098.559

(d) Language identification/ diarization Output: LID_< evaluation_id >.csv

Audio File Name, speaker, confidence score (in %), start TS, end TS

Example:

ps6_01_001.wav, english, 95, 100.006, 120.002

ps6_01_001.wav, hindi, 93, 118.080, 200.256

ps6_01_001.wav, english, 90, 234.130, 339.786

ps6_01_002.wav, hindi, 90, 234.130, 339.786

ps6_01_003.wav, hindi, 85, 014.540, 058.954

ps6_01_003.wav, punjabi, 75, 060.240, 098.559

(e) ASR Output: ASR_< evaluation_id >.trn

Audio File Name, start TS, end TS, Transcript

Example:

ps6_01_001.wav, 100.006, 120.002, hi how are you?

ps6_01_001.wav, 118.080, 200.256, आज थोड़ा काम है

ps6_01_001.wav, 234.130, 339.786, come here please

ps6_01_002.wav, 234.130, 339.786, क्या चल रहा है?

ps6_01_003.wav, 014.540, 058.954, आप कैसे हैं?

ps6_01_003.wav, 060.240, 098.559, ਤੁਸੀਂ ਕਿਵੇਂ ਹੋ?

(f) NMT Output: NMT_< evaluation_id >.txt

Audio File Name, start TS, end TS, Translation

Example:

ps6_01_001.wav, 100.006, 120.002, hi how are you?

ps6_01_001.wav, 118.080, 200.256, I have some work today

ps6_01_001.wav, 234.130, 339.786, come here please

ps6_01_002.wav, 234.130, 339.786, What's happening?

ps6_01_003.wav, 014.540, 058.954, How are you doing?

ps6_01_003.wav, 060.240, 098.559, How are you doing?

5. **Solution Hash.** Pack your complete solution used for generating the above results in a single compressed file and generate the hash. Write the hash in text file with name PS_06_<application_id/registration_id>_<evaluation_id>.hash. **Please note that the participant may be asked to use this compressed file (docker/ virtual file) along with hash to verify their shortlisting\evaluation results at any stage of Grand Challenge**.

6. The result data (para 3 & 4 above), hash file (para 5 above) along with solution description should be submitted in a compressed file with name PS_06_

<application_id/registration_id>_<evaluation_id>.<ext>. The solution description should include the following:

(a) Solution Brief:

(b) Approach & Architecture of Solution:

(c) Machine learning frameworks used (e.g., PyTorch, Tensorflow, CNTK):

(d) Training Data Set used:

(e) Average time taken in generation of results:

(f) Hardware Requirements:

(i) Total number of CPU cores used

(ii) Description of CPUs used (model, speed, number of cores)

(iii) Total number of GPUs used

(iv) Description of GPUs used (model, single precision TFLOPS, memory)

(v) Total number of TPUs used

(vi) Generations of TPUs used (e.g., v2 vs v3)

(vii) Total available RAM

(viii) Used disk storage

(ix) Any other details

Appendix 'B'

Evaluation Formulas

The evaluation formulas for each of the metrics are as given below:

(a) Speaker Identification (Weightage 10%):

Top-1 Identification Accuracy = Number of correctly identified speakers / Total number of utterances × 100

(b) Speaker diarization (Weightage 15%):

DER (Diarization Error Rate) = (False Alarm Duration + Missed Speech Duration + Speaker Error Duration) / Total Reference Speech Duration

False Alarm: Time when the system detects speech when there is none, or assigns a speaker when no one is talking.
Missed Speech: Time when speech occurs but is not detected.
Speaker Error: Time when the speaker label is wrong.
Total Reference duration: Total speech duration (summation of all the speakers segments' duration).

(c) Language identification/ diarization (Weightage 15%):

DER (Diarization Error Rate) = (False Alarm Duration + Missed Language Duration + Language Error Duration) / Total Reference Duration

False Alarm: Time when the system detects speech when there is none, or assigns a language when no one is talking.
Missed Language: Time when speech occurs but no language is detected.
Language Error: Time when the language label is wrong.
Total Reference duration: Total speech duration (summation of all the speakers segments' duration).

(d) Automatic Speech Recognition (Weightage 20%):

WER (Word Error Rate) = (Substitutions + Deletions + Insertions) / Total Number of Reference Words

Insertion: a word is added that wasn't said in the transcript.
Deletion: a word is not added that was said in the transcript.
Substitution: a word is replaced with a different word to the transcript.

(e) Neural Machine Translation (Weightage 10%):

BLEU = Brevity Penalty × exp(average log n-gram precision)

Brevity Penalty = e^ (1 - R/C)

Where: n = 1 to 4
R : Length of Reference Sentence
C : Length of translated Sentence