

The ALBAYZIN 2026 Speech-COSER Evaluation Plan

Javier Tejedor¹, Laura Herrera², Alicia Lozano-Diez², Clara Adsuar Ávila³,
Doroteo T. Toledano², Inés Fernández-Ordóñez³

¹ Escuela Politécnica Superior, Universidad San Pablo-CEU, CEU Universities,
Madrid, Spain.

`javier.tejedornoguerales@ceu.es`

² AUDIAS - Audio, Data Intelligence and Speech, Escuela Politécnica Superior,
Universidad Autónoma de Madrid, Madrid, Spain

`laura.herrera@uam.es`

`alicia.lozano@uam.es`

`doroteo.torre@uam.es`

³ Departamento de Filología Española, Universidad Autónoma de Madrid, Madrid,
Spain

`clara.adsuar@uam.es`

`ines.fernandez-ordonez@uam.es`

Abstract. This document presents the evaluation plan for the upcoming ALBAYZIN 2026 Speech-COSER evaluation. Three different tracks are proposed: Automatic Speech Recognition (ASR), Speaker Diarization (SD) and Spoken Term Detection (STD). These tracks are evaluated on the Spanish rural speech corpus COSER, which consists of different interviews carried out in Spanish rural areas. The ASR track consists of identifying the words that are pronounced in the speech content. The SD track consists of identifying the speaker segments in each speech file. Finally, the STD track consists of finding a list of terms within the audio files. The challenge provides training/development data for system construction and test data for system evaluation. Standard metrics such as Word Error Rate (WER), Diarization Error Rate (DER) and Actual Term-Weighted Value (ATWV) will be used for ASR, SD and STD system evaluation, respectively. The evaluation will also integrate a publicly available leaderboard for system submission and performance evaluation.

1 Introduction

The *ALBAYZIN 2026 Speech Technologies for COSER: A Spanish Rural Corpus (Speech-COSER)* challenge is supported by the Spanish Thematic Network on Speech Technology (RTTH) and is organized by Universidad San Pablo-CEU and AUDIAS from Universidad Autónoma de Madrid.

The challenge focuses on several speech technologies on the Spanish speech rural corpus called COSER ⁴. This Spanish dataset stands out due to its substantial geographical and dialectal diversity, covering all regions in Spain, and

⁴ <https://corpusrural.es/ING/index.php>

whose speakers belong to an older population in rural areas with low education levels. This is the largest corpus of dialectal European Spanish, with recordings of rural contexts and older Spanish speakers. The complexity of this dataset presents a significant challenge for speech technologies.

The challenge integrates three different tracks: Automatic Speech Recognition (ASR), Speaker Diarization (SD) and Spoken Term Detection (STD).

2 Evaluation description

The evaluation consists of three different tracks:

- Automatic Speech Recognition (ASR): This track focuses on automatically transcribing the audio files, for which the sequence of words that appears in each audio file must be provided.
- Speaker Diarization (SD): This track focuses on automatically segmenting the audio files according to speaker turns, for which timestamps and speaker assignment to segments must be provided. Even if the identity of the speakers and number of speakers are unknown, the output should group segments of the same speaker under the same speaker label.
- Spoken Term Detection (STD): This track aims to detecting a list of terms within the audio files. This list is assumed to be unknown when processing the audio. A set of occurrences for each term detected in the audio files must be generated, along with their timestamps and score as output.

3 Database description

The COSER corpus [1] was specifically created to capture Spanish varieties spoken in different rural regions and it is the largest corpus of European Spanish dialects. It includes interviews and informal conversations with speakers from rural areas of Spain, whose informants generally have low levels of education and an average age of 74.2 years. On the other hand, Spanish shows notable regional variation, and in certain areas, other languages, such as Galician, Catalan or Basque, are also spoken. Consequently, some of the recordings in the COSER corpus include interviews conducted in both Spanish and these languages, further enriching the diversity of linguistic data. Moreover, the corpus presents a complex environment due to the diverse recording conditions, background noise, and the presence of different dialectal varieties.

The audios consist of semi-guided interviews that aim to cover topics of traditional life, but informants are free to change the conversation to other topics. There are 22 defined and labeled topics that are included in the metadata of each transcription. The recordings were obtained during more than 35 years, from 1990 to April 2025, in 1468 rural locations (57 provinces or islands). In total, the dataset consists of 2002 hours of audio, of which only the transcription of about 350 hours is available. Interviews have an average length of 1 hour and 5 minutes but the duration of each varies from 30 minutes to 3 hours. The dataset

has an overall balance in gender representation, but it does not ensure balance across different locations. Regarding the number of speakers in each audio, only one person is interviewed in detail, but interruptions by other individuals are not avoided. Since the interviews have been conducted from a linguistic point of view, the transcripts include morphosyntactic tags (available only in XML version in the download area⁵). Therefore, the textual data of the interviews include metadata with general information about the audio and the full transcript. The metadata include the province and location where the recording was made, the date and duration, information about the speakers and interviewers, and the main topics discussed. The transcriptions are highly detailed, as they contain a segmentation by speaker turn (including speaker tags), conversation marks and the textual transcription, including punctuation and some orthographic adaptations to dialectal phonetic features. Time alignment is by speaker turns rather than by individual words.

Therefore, this corpus is an exceptional resource due to its extensive representation of the linguistic varieties present in Spain, which are often under-represented in ASR/SD/STD models, particularly for elderly people with low education levels living in rural areas, highlighting the need for effective adaptation to these contexts.

The COSER speech data were originally recorded in several audio formats (e.g., mono and stereo) and different sampling frequency. All the speech data were converted to PCM, 16khz, single channel and 16 bits per sample WAV files using the SoX tool for this evaluation. Moreover, the textual data in this database were also adapted to the required format of the different tracks presented in this evaluation.

For this challenge, we have selected part of the corpus to create our training/development and test datasets, which have been further curated for the addressed tasks.

3.1 Training/development data

The training and development data provided to participants can be used in any way (i.e., training, development, etc.) for system construction. In case participants use the COSER corpus for training/development, only the audio files included in the provided list may be used. The use of any other audio from the corpus is not allowed.

For the **ASR track**, approximately 23 hours of audio segments, with a maximum duration of 30 seconds each, will be provided with their manually validated transcriptions as training/development data. In addition, 10 hours of full audio recordings will be available as training/development data. However, the transcriptions associated with these full-length audios have not been cleaned nor manually verified and may contain errors. The scoring script will be based on the *meeteval* toolkit [3]. This script, along with the necessary input files, will be provided to participants to ensure the reproducibility of the results.

⁵ <https://corpusrural.fe.uam.es/coser/descargas.php>

For the **SD track**, the same set of audio segments and full-length recordings as for the ASR track will be provided. For this task, the corresponding RTTM files will also be made available. Some segments may contain multiple speakers, including overlapping speech. Note that the labels of the full-length recordings may also contain errors. The scoring tool will be based on the official *dscore* tool [4]. The corresponding evaluation script and the necessary input files will also be provided to participants.

For the **STD track**, about 1 hour of speech will be provided to participants. Specifically, 115 audio files with about 30 seconds each will be provided as training/development speech data. Regarding the list of terms for search in the training/development data, this consists of 121 terms whose length ranges from 4 to 22 single graphemes. A term can be composed by one or more words and may appear or not in the audio files. In case the term consists of more than one word, the symbol `_` is used as word boundary. The orthographic transcriptions of the selected list of terms along with the occurrences and timestamps for each of these terms in the training/development speech data will be provided. The scoring tool along with the necessary input files to run it for training/development data will also be provided. The systems will be scored using the NIST STD scoring tool [2].

3.2 Test data

For the **ASR track**, 30 hours of audio will be provided, comprising 22 recordings with durations between 50 minutes and 2.5 hours.

For the **SD track**, the same recordings as for the ASR track will be provided for speaker diarization.

For the **STD track**, the test speech data total about 1.7 hours. Specifically, 27 audio files that range between 3 and 5 minutes will be provided as test speech data. Regarding the list of terms for search in the test data, this consists of 321 terms whose length ranges from 3 to 18 single graphemes. A term can be composed by one or more words and may appear or not in the audio files. In case the term consists of more than one word, the symbol `_` is used as word boundary. The orthographic transcriptions of the selected list of terms will be provided to participants.

None of these test data are allowed to be used as training or development data in any form. Moreover, no human analysis, supervision, revision, or labeling of the test data is permitted.

4 Evaluation of system performance

4.1 ASR track

For the ASR track, the Word Error Rate (WER) metric will be used for system evaluation. This metric will be calculated as the sum of substitutions, deletions and insertions divided by the total number of words in the reference transcription. The evaluation will be conducted on both the transcriptions in their original

form (which include punctuation marks) and on normalized transcriptions, with all text converted to lowercase and containing no punctuation marks. Non-speech events such as laughter, verbal nods, and breaths, will not be considered in the evaluation.

4.2 SD track

For the SD track, the Diarization Error Rate (DER) will be the metric used for system evaluation. The DER metric will be computed based on the time associated with false alarms, missed speech, and speaker confusion errors. Overlapping speech segments, where multiple speakers are active simultaneously, will also be included in the evaluation. No collar will be applied, meaning that no temporal tolerance will be allowed at the segment boundaries.

4.3 STD track

The Actual Term Weighted Value (ATWV) [2] will be the primary metric for the STD track. Participants will be ranked from the ATWV obtained on the test data. Maximum Term Weighted Value (MTWV) scores and Detection Error Tradeoff (DET) [2] curves for the STD track will also be computed as secondary metrics for system analysis.

5 General evaluation conditions

5.1 Data download

Instructions for downloading will be given to registered participants at due time for the training/development and test data.

5.2 System output format and submission

ASR track. For the ASR track, results must be submitted as a ZIP file containing, for each audio file, a full-audio transcription with punctuation marks in a plain text file (e.g., *Nosotros teníamos al lao del caserío ese, había una..., casa que vivían unos pastores. Tenían ovejas, tienen también, pero ahora menos, muchas ovejas.*) Each file should be named `<record_id>_fullaudio_transcrip.txt`. The `<record_id>` must match the corresponding WAV file `<record_id>.wav`. The ZIP file should be named `<ID_group>_ASR_submission.zip`.

SD track. For the SD track, results must be also submitted as a ZIP file containing, for each audio file, a Rich Transcription Time Marked (RTTM) file with the speaker diarization output. The RTTM files must follow the standard format:

```
SPEAKER <record_id> onset duration <NA> <NA> spk_id <NA> <NA>.
```

Each RTTM file should be named `<record_id>.rttm`, and the ZIP file as `<ID_group>_SD_submission.zip`.

STD track. For the STD track, detection results must be sent in a single xml file according to the ‘stdlist’ XML format specified in the NIST STD 2006 evaluation plan [2]. Please note that for this track, timestamps for each detection are relevant, since they are taken into account by the NIST STD scoring tool to evaluate if each term detection is correct or not. Higher scores mean more confidence in the detection appearing in the corresponding speech file between the given timestamps. The xml file should be included in a ZIP file named `<ID_group>_STD_submission.zip`.

For all the tracks, the evaluation will be performed in a continuous evaluation method, with a publicly available leaderboard where participants will be able to submit their output files for each of the tracks on the test data and check their performance with respect to other participants. This leaderboard will be frozen by the evaluations submission deadline and will be reopened for post-evaluation period. More information regarding this leaderboard will be released to registered participants.

Participants can submit their systems for any track and the amount of submissions is limited to 2 per track and day.

Participants will be ranked in each track according to the performance attained by the best submitted system on the test data.

5.3 Registration and system description

Registration rules. Interested groups must register for the evaluation before July 31st, 2026 sending an email to javier.tejedornoguerales@ceu.es and alicia.lozano@uam.es with the name group, acronym name (which will be the `< ID_group >` with which participants need to submit their systems), and information on the tracks the group aims to register.

System description. Research groups must provide a file with the description of the submitted systems in which the results obtained in, at least, the training/development data, must be included.

Participants can choose between two submission ways:

- The first way relies on editing the system description paper following the IberSPEECH 2026 paper submission template so that the submitted paper (describing the system/s and the results) will appear in the IberSPEECH 2026 proceedings following the regular peer review process (deadline set at 22nd June). Moreover, participants may also have the chance to submit an extended version of this paper to a journal. This submission way implies sending one or more representatives to the evaluation workshop, to be held in Madrid, Spain as part of IberSPEECH 2026 (November 2026), and present there upon acceptance of the paper.
- The second way demands a free-format document in which participants describe the submitted system/s along with the results, but this will not appear

in the IberSPEECH 2026 proceedings. In this case, participants are allowed to present on-line their system/s without physically attending the conference, or send a video to the evaluation organizers explaining their submitted system/s, which will be shown during the evaluation workshop. The paper submission deadline in the second submission way is September 30th, 2026 (23:59 GMT+1).

5.4 Schedule

- April 20th, 2026. Registration opens.
- May 1st, 2026. Release of the training/development data.
- June 5th, 2026. Release of the test data. System submission (leaderboard) opens.
- July 31st, 2026. Registration deadline.
- September 30th, 2026 (23:59, GMT +1). System submission and system description paper deadline (for system description paper according to the second system description submission way). The leaderboard will be frozen by this date.
- October 31st, 2026. Results are distributed to the participants. The leaderboard will be reopened by this date for post-evaluation period.
- November 18th-20th, 2026. IberSPEECH 2026 Albayzin Evaluations special session in Madrid.

6 Additional information and summary of evaluation rules

- Interested groups must register for the evaluation before July 31st, 2026 sending an email to *javier.tejedornoguerales@ceu.es* and *alicia.lozano@uam.es* with the name of the group, its acronym and the list of tracks to register.
- Starting from May 1st, 2026, and once registration data are validated, the training/development data will be released only to registered participants.
- The test data will be released by June 5th, 2026. The paper submission deadline in the second submission way is September 30th, 2026 (23:59 GMT+1).
- Registered groups commit themselves to use the provided data only for research purposes. Authors are also requested to cite the Albayzin 2026 Speech-COSER system description paper that will be included in the IberSPEECH 2026 Proceedings (if authors choose the first submission way for system description) and the COSER corpus in case of using the COSER data in future publications.
- No manual intervention is allowed for each system developed to generate the final output file and hence, all the developed systems must be fully automatic. Listening to the test data, or any other human interaction with the test data is forbidden before all the results have been submitted.

- In case the participant site has submitted a paper to appear in the Iber-SPEECH proceedings, it is mandatory to send one or more representatives to the evaluation workshop, to be held in Madrid, Spain as part of Iber-SPEECH 2026 (November 2026). However, in case the participant site has just submitted a free-format document with the system description+results, it is allowed to present on-line their system without physically attending the conference, or send a video to the evaluation organizers explaining their submitted systems, which will be shown during the evaluation workshop.
- This plan might be modified due to new restrictions or unplanned needs, to detected errors or inaccuracies. Updated versions of this plan, if any, will be announced through the Speech-COSER evaluation website and emailed to the registered participants.

7 Acknowledgements

This research was supported by project PID2021-125943OB-I00 funded by MCIN/AEI/10.13039/501100011033/FEDER, UE, project PID2024-160789OB-I00 funded by MICIU/AEI/10.13039/501100011033/FEDER, UE, project TSI-100927-2023-0003/TSI-100927-2023-0004/TSI-100927-2023-0005 funded by the “Ministerio para la transformación digital y la función pública”, project SI4/PJI/2024-00237 (COSER-IA), funded by Comunidad de Madrid and project PREP2024-003414 funded by MICIU/AEI/10.13039/501100011033 and FSE+.

Computational resources for this research were partly provided by CCC-UAM.

References

1. Fernández-Ordóñez, I.: COSER (Corpus Oral y Sonoro del Español Rural) (2005), <https://corpusrural.es/ING/index.php>
2. Fiscus, J.G., Ajot, J.G., Garofolo, J.S., Doddington, G.: Results of the 2006 spoken term detection evaluation. In: Proc. of ACM SIGIR. pp. 1–4 (2007)
3. von Neumann, T., Boeddeker, C., Delcroix, M., Haeb-Umbach, R.: MeetEval: A toolkit for computation of word error rates for meeting transcription systems. In: Proc. 7th International Workshop on Speech Processing in Everyday Environments (CHiME 2023). pp. 27–32 (2023)
4. Ryant, N.: dscore: Nist diarization error rate scoring script (2020), <https://github.com/nryant/dscore>