Albayzin Evaluation
IberSPEECH-RTVE 2024 Speaker Diarization and Identity Assignment

Alfonso Ortega[1], Antonio Miguel[1], Eduardo Lleida[1], Virginia Bazán[2], Carmen Pérez[2], and Pere Vila[2]

[1] Vivolab, Aragon Institute for Engineering Resarch (I3A)
University of Zaragoza, Spain
{ortega,amiguel,lleida}@unizar.es
http://www.vivolab.es
[2] Corporación Radiotelevisión Española, Spain http:www.rtve.es

SD and ID-A evaluation plan v1 (May 17, 2024)

Abstract. Supported by the *Spanish Thematic Network on Speech Technology* (RTTH)[3] and *Cátedra RTVE Universidad de Zaragoza*[4] and organized by ViVoLab Universidad de Zaragoza, the new edition of the IberSPEECH-RTVE Speaker Diarization challenge is a new event in the ALBAYZIN evaluation series.The evaluation will be conducted as part of the Iberspeech 2024[5] conference to be held in Aveiro, Spain from 11 to 13 November 2024.

The task of grouping all the turns of the same speaker in an audio document under the same label is usually known as speaker diarization. In addition to this, if there exists prior knowledge of the identity of the people involved, Speaker Identification technologies can be used to assign a name to each diariziation label. For applications such as spoken document retrieval or audio indexing this is a crucial task and for some other speech technologies like Automatic Speech Recognition or Speaker Identification can be regarded as a supporting technology.

In this evaluation, broadcast audio documents will be considered with limited prior knowledge about the identity of the speakers involved in the audio under test and no prior knowledge about their number. This year, two tasks will be proposed: First a classical speaker diarization task in which participants must provide the labels for the segments in which different speakers appear assigning the same label for all the utterances of the same speaker. Second, on top of the diarization labels, participants are required to assign specific identities for a limited number of participants. For those speakers, several utterances will be provided in order to allow the system to build speaker models.

*Speaker diarization is **mandatory** in order to participate in the challenge. Identity assignment is **optional**, but we encourage participants to give a try.*

[3] http://www.rthabla.es
[4] http://catedrartve.unizar.es
[5] http://iberspeech.tech

The Diarization Error Rate will be used as scoring metric as defined in the RT evaluations organized by NIST.
Regarding the allowed training material, only one condition is proposed this year, an open-set condition in which not only the data provided within this Albayzin evaluation but external data can also be used for training as long as they are publicly accessible to everyone (not necessarily free).

1 Introduction

Some tasks such as Spoken document retrieval (SDR), Automatic Speech Recognition (ASR) systems for Broadcast shows, Speaker Identification (SPK ID) or multimedia indexing (MI) in very large repositories, Speaker Diarization and Identity assignment are considered very important tasks. Therefore, the development of accurate Speaker Diarization and Identity Assignment systems is essential to allow applications like SDR, ASR, SPK ID or MI to perform adequately in real-world environments.

The Speaker Diarization and Identity Assignment evaluation consists of segmenting broadcast audio documents according to different speakers and linking those segments which originate from the same speaker. On top of that, for a limited number of speakers, assign the name of these people to the correct diarization labels.

2 Database description and partitions

As in previous editions, the evaluation database, RTVE2024DB, has been donated by Corporación Radiotelevisión Española (RTVE) and labeled thanks to the *Cátedra RTVE de la Universidad de Zaragoza*. Also, the RTVE datasets, RTVE2018DB [1], RTVE20220DB and RTVE2022DB proposed for the 2018, 2020 and 2022 Albayzin Evaluations, [6], the Catalan broadcast news database from the 3/24 TV channel proposed for the 2010 Albayzin Audio Segmentation Evaluation [2, 3] and the Corporación Aragonesa de Radio y Televisión (CARTV) database proposed for the 2016 Albayzin Speaker Diarization evaluation will be provided if needed for training or development purposes.

2.1 Training and Development data

RTVE2018 database. For training and development purposes, the RTVE2018 database contains one training partition, two development partitions and a test partition. This database can be used for both training and development. Around **37** hours with diarization and reference speech segmentation are included and can be used for any purpose including system development or training In addition to this, for a limited number of speakers, several audio excerpts are provided in

[6] `http://http://catedrartve.unizar.es/rtvedatabase.html`

order to allow participants to build speaker models in the Identity Assignment task.

The RTVE2018 database development partitions correspond to two different debate shows, four programs (7:26 hours) of *La noche en 24H*[7], where a group of political analysts comment what happened throughout the day, and eight programs (7:42 hours) of *Millenium*[8] where a group of experts debates about a current issue. Regarding the eval set, this partition consists of 22:45 h of four different shows, 22 episodes of "España en comunidad", which corresponds to 8:09 hours, 8 episodes of "Latinoamerica en 24H" with 4:07 hours, 1 episode of "La Mañana" with 1:36 hours, and 9 episodes of "La Tarde en 24H", which corresponds to 8:52 hours. The data is distributed in AAC format, (LC mp4a), 44100 Hz, stereo, variable bitrate.[9]

RTVE2020 database The RTVE2020 database contains a small development partition with around 4 hours of audio labeled with speaker turns is included in the database and test partition with 29 hours of different shows labeled with speaker turns and more than 150 speaker identities with their enrollment audio data.

RTVE2022 database The RTVE2022 database [6] contains a test partition with 25 hours of different shows labeled with speakers turns and 74 speaker identities with their enrollment audio data. As enrollment for each speaker, an audio recording of at least 30 s is provided.

More information about the content of the dataset can be found in `https://catedrartve.unizar.es/rtvedatabase.html`

Aragón Radio database. The database donated by the Corporación Aragonesa de Radio y Televisión (CARTV) consists of around twenty hours of the Arag´on Radio broadcast. This data set contains around 85% of speech, 62% of music and 30% of noise in a way that 35% of the audio contains music along with speech, 13% is noise along with speech and 22% is speech alone. The data will be supplied in PCM format, mono, little endian 16 bit resolution, and 16 kHz sampling frequency.

3/24 TV channel database. The Catalan broadcast news database from the 3/24 TV channel proposed for the 2010 Albayzin Audio Segmentation Evaluation [2, 3] was recorded by the TALP Research Center from the UPC in 2009 under the Tecnoparla project [4] funded by the Generalitat de Catalunya. The Corporació Catalana de Mitjans Audiovisuals (CCMA), owner of the multimedia content,

[7] http://www.rtve.es/alacarta/videos/la-noche-en-24-horas/
[8] http://www.rtve.es/alacarta/videos/millennium/
[9] We recommend ffmpeg to change to your audio format
https://www.ffmpeg.org/
ffmpeg -i file.aac -ar 16000 -ac 1 file.wav

allows its use for technology research and development. The database consists of around 87 hours of recordings in which speech can be found in a 92% of the segments, music is present a 20% of the time and noise in the background a 40%. Another class called *others* was defined which can be found a 3% of the time. Regarding the overlapped classes, 40% of the time speech can be found along with noise and 15% of the time speech along with music. The data will be supplied in PCM format, mono, little endian 16 bit resolution, and 16 kHz sampling frequency.

2.2 Evaluation data

RTVE2022 database. The evaluation data will contain a set of TV shows covering a variety of scenarios from RTVE2022 database. The *test* partition contains around 25 h of audiovisual documents labeled in terms of speaker turns. No a priory knowledge will be provided about the number of the speakers participating in the audio to be analyzed.

More than 100 characters have been labeled and their corresponding enrollment files needed for speaker identification are provided. The enrollment material consists of one or more audio files with more than 30 seconds of speech of each known character.

The detailed information about the evaluation data will be released by September 2nd coinciding with the beginning of the evaluation task. The data will be distributed in AAC format, (LC mp4a), 44100 Hz, stereo, variable bitrate.[10]

3 Diarization Scoring

As in the NIST RT Diarization evaluations [5], to measure the performance of the proposed systems, the Diarization Error Rate (DER) will be computed as the fraction of speaker time that is not correctly attributed to that specific speaker. This score will be computed over the entire file to be processed; including regions where more than one speaker is present (overlap regions).

This score will be defined as the ratio of the overall diarization error time to the sum of the durations of the segments that are assigned to each class in the file.

Given the dataset to evaluate Ω, each document is divided into contiguous segments at all speaker change points found in both the reference and the hypothesis, and the diarization error time for each segment n is defined as

$$E(n) = T(n)\left[\max\left(N_{ref}(n), N_{sys}(n)\right) - N_{Correct}(n)\right] \tag{1}$$

where $T(n)$ is the duration of segment n, $N_{ref}(n)$ is the number of speakers that are present in segment n, $N_{sys}(n)$ is the number of system speakers that

[10] We recommend ffmpeg to change to your audio format https://www.ffmpeg.org/

are present in segment n and $N_{Correct}(n)$ is the number of reference speakers in segment n correctly assigned by the diarization system.

$$DER = \frac{\sum_{n \in \Omega} E(n)}{\sum_{n \in \Omega} (T(n) N_{ref}(n))} \tag{2}$$

The diarization error time includes the time that is assigned to the wrong speaker, missed speech time and false alarm speech time:

- **Speaker Error Time**: The Speaker Error Time is the amount of time that has been assigned to an incorrect speaker. This error can occur in segments where the number of system speakers is greater than the number of reference speakers, but also in segments where the number of system speakers is lower than the number of reference speakers whenever the number of system speakers and the number of reference speakers are greater than zero.
- **Missed Speech Time**: The Missed Speech Time refers to the amount of time that speech is present but not labeled by the diarizaton system in segments where the number of system speakers is lower than the number of reference speakers.
- **False Alarm Time**: The False Alarm Time is the amount of time that a speaker has been labeled by the diarization system but is not present in segments where the number of system speakers is greater than the number of reference speakers.

Consecutive segments of the same speaker with a silent of less that 2 seconds come together and are considered as a single segment. A forgiveness collar of ± 0.25 seconds, before and after each reference boundary, will be considered in order to take into account both inconsistent human annotations and the uncertainty about when a speaker begins or ends.

4 Identity Assignment Scoring

For the Identity Assignment Task, **Assignment Error Rate (AER)** will be used which is a slightly modified version of the previously described Diarization Error Rate. This metric is defined as the amount of time incorrectly attributed to the speakers of interest divided by the total amount of time that those specific speakers are active. Mathematically it can be expressed as

$$AER = \frac{FA + MISS + SPEAKER\ ERROR}{REFERENCE\ LENGTH} \tag{3}$$

where

- **FA**: Represents the False Alarm Time which contains the length of the silence segments o speech segments that belong to unknown speakers incorrectly attributed to a certain speaker.

– **MISS**: Represents the Missed Speech Time which takes into account the length of the speech segments that belong to speakers of interest not attributed to any speaker.
– **SPEAKER ERROR**: The Speaker Error Time considers the length of the speech segments that belong to speakers of interest attributed to an incorrect speaker.
– **REFERENCE LENGTH**: The reference length is the sum of the lengths of all the speech segments uttered by the people of interest. Those identities for which the participants will have audio to train their models.

For certain tasks such as Multimedia Indexing for example, it is very important to assign correctly as much of the content as possible but it is also essential not to miss any speaker even though the amount of time they appear in the document is small. Therefore, as an alternative metric, we propose for this evaluation the use of the **Average Speaker Error (ASE)** over all the speakers of interest, that can be defined as:

Let N be the number of speakers of interest, the average speaker error can be obtained as

$$ASE = \frac{1}{N} \sum_{i=1}^{N} \frac{dur(miss_i) + dur(fa_i)}{dur(ref_i)} \tag{4}$$

where $dur(miss_i)$ is the total duration of all miss errors for the ith speaker of interest, $dur(fa_i)$ is the total duration of all false alarm errors for the ith speaker of interest, and $dur(ref_i)$ is the total amount of time the ith speaker of interest is active according to the reference. Using this metric an incorrectly assigned segment computes as a miss error for a speaker of interest and a false alarm error for another. Due to the fact that the speaker's activity distribution is clearly unbalanced, errors from different speakers of interest are weighted differently according to the total amount of time a given speaker of interest is active in the database.

In both metrics, AER and ASE, a collar of ± 0.25 seconds around each reference boundary is not scored in order to avoid uncertainty about when an acoustic class ends or begins, and to consider inconsistent human annotations.

4.1 Segmentation Scoring Tool and Speaker Diarization System Output Files

The tool used to obtain the Diarization Error Rate (DER) and the Assignment Error Rate (AER) is the one developed for the RT Diarization evaluations by NIST "md-eval-v22.pl", and for the Average Speaker Error (ASE) the tool is "ase-eval-v1.py´´ both available at the score folder of the RTVE2024DB distribution.

The format's definition for the submission of the Speaker Diarization results has been fixed according to the operation of the NIST's tool. Specifically the Rich Transcription Time Marked (RTTM) format will be used for speaker diarization

and identity assignment system outputs and reference files. RTTM files are space-separated text files that contain meta-data 'Objects' that annotate elements of each recording and a detailed description of the format can be found in Appendix A of the 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan [5]. Thus, the required information for each segment will be:
SPEAKER File Channel Beg_Time Dur <NA> <NA> Speaker_Label <NA> <NA>
where:

- **SPEAKER**: A tag indicating that the segments contain information about the beginning, duration, identity, etc. of a segment that belongs to a certain speaker.
- **File**: It is the name of the considered file.
- **Channel**: It refers to the channel. Since we are dealing with mono recordings this value will always be 1.
- **Beginning_Time**: The beginning time of the segment, in seconds, measured from the start time of the file.
- **Duration**: It indicates the duration of the segment, in seconds.
- **Speaker_Label**: It refers to the label asigned to the speaker present in the considered segment.

The tag <NA> indicates that the rest of the fields are not used. The numerical representation must be in seconds and hundredth of a second. The decimal delimiter must be '.'.

For the Albayzin 2024 Speaker Diarizaton evaluation the best mapping option between hypothesis and reference labels will be used. The command line is:

md-eval-v22.pl -c 0.25 -b -r <SPKR-REF>.rttm -s <SYSTEM>.rttm

Unlike previous use, for the Albayzin 2024 Identity Assignment evaluation in order to obtain the AER the best mapping option between hypothesis and reference labels will NOT be used. The command line is:

md-eval-v22.pl -c 0.25 -r <SPKR-REF>.rttm -s <SYSTEM>.rttm

To obtain the Average Speaker Error (ASE) the command line is:

python3 ase-eval-v1.3.py reference.rttm hypotesis.rttm

5 General Evaluation Conditions

The organizers encourage the participation of all researchers interested in speaker diarization. All teams willing to participate in this evaluation must registered through the challenge web page
`http://catedrartve.unizar.es/albayzin2022.html`
before September 2nd, 2024.
In case of any difficulty, you can send an e-mail to lleida@unizar.es and ortega@unizar.es with CC to Iberspeech 2024 Evaluation organizers at albayzinevaluations@gmail.com

5.1 Data License Agreement

The RTVE data is available to the evaluation participants and subject to the terms of a licence agreement with the RTVE. The license agreement can be downloaded from Cátedra RTVE-UZ web page:
`http://catedrartve.unizar.es/rtvedatabase.html`

Participants must sign the agreement (digital signatures is valid) and send a copy attached to the email. A copy signed by RTVE representative will be returned. Please read carefully the information provided on the Cátedra RTVE-UZ web page related with the use of the RTVE data after the evaluation campaign.

5.2 Evaluation Rules

Each participant team must submit at least a primary system, but they can also submit up to two contrastive systems. Each and every submitted system must be applied to the whole test database. The ranking of the evaluation will be done according to results of the primary systems but the analysis of the results of the contrastive systems will be also processed and presented during the evaluation session at Iberspeech. All participant sites must agree to make their submissions (system output, system description, ...) available for experimental use by the rest of the participants and the organizing team.

The participant teams will notify and provide the total time required to run the set of tests for each submitted system (specifying the computational resources used). No manual intervention is allowed for each developed system to generate its output, thus, all developed systems must be fully automatic. Listening to the evaluation data, or any other human interaction with the evaluation data, is not allowed before all results have been submitted. The evaluated systems must use only audio signals. Any publicly available data can be used for training together with the data provided by the organization team to train the speaker diarization system only in the open-set condition. In case of using additional material, the participant will notify it and provide the references of this material. These databases must be publicly accessible although not necessarily free.

5.3 Result Submission Guidelines

The evaluation results must be presented in just one ZIP file per submitted system. The ZIP file must contain one RTTM file per submitted system. The file output file must be identified by the following code:

EXP-ID::=<SITE>_<SYSID> where,

- <**SITE**>: Refers to a three letter acronym identifying the participant team (UPM, UPC, UVI, ...).
- <**SYSID**>: Is an alphanumeric string identifying the submitted system. For the primary system the SYSID string must begin with p-, c1- for contrastive system 1 and c2- for contrastive system 2.

Each participant team must upload the zip files through the challenge web page
`http://catedrartve.unizar.es/albayzin2024.html`

In case of uploading problems send an e-mail with the corresponding ZIP result files to

- lleida@unizar.es
- ortega@unizar.es

5.4 System Descriptions

Participants must send, along with the result files, a PDF file with the description of each submitted system. The format of the submitted documents must fulfil the requirements given in the IberSpeech 2024 call for papers. You can use the templates provided for the Iberspeech conference (WORD or LaTeX). Please, include in your descriptions all the essential information to allow readers to understand the key aspects of your systems.

A full conference paper can be submitted to the IberSpeech Conference as a regular paper for the Albayzin Evaluation special session. Please, take advise of the deadlines in the IberSpeech 2024 web page
`https://iberspeech.tech/`

6 Schedule

- May 20th, 2024: Registration opens and release of the training data.
- June 3rd, 2024: Release of training and development data
- July 31st, 2024: Registration deadline
- September 2nd, 2024: Release of the evaluation data.
- October 18th, 2024: Deadline for submission of results and system descriptions.
- October 31tst, 2024: Results distributed to the participants.
- November 12th, 2024: Official results presented publicly and published
- November 12th, 2024: IberSpeech 2024 special session in Aveiro.

7 Acknowledgments

The organizing team would like to thank to Corporación Radiotelevisión Española and Cátedra RTVE de la Universidad de Zaragoza by their efford for providing the data for the 2020 evaluation. Also, the Corporación Aragonesa de Radio y Televisión and Aragón Radio for providing the additional data for the evaluation. Thanks also to Martin Zelenak and Javier Hernando who organized the 2010 Albayzin Audio Segmentation Evaluation for their help, support and for providing the training material for this evaluation. And also to the organizing committee of Iberspeech 2024 for their help and support.

References

[1] Lleida, E.; Ortega, A.; Miguel, A.; Bazán-Gil, V.; Pérez, C.; Gómez, M.; de Prada, A. Albayzin 2018 Evaluation: The IberSpeech-RTVE Challenge on Speech Technologies for Spanish Broadcast Media. Appl. Sci. 2019, 9, 5412.

[2] Zelenak, M., Schulz, Hernando, J., Albayzin 2010 Evaluation Campaign: Speaker Diarization. VI Jorandas en Tecnologías del Habla, FALA 2010. Vigo, Noviembre 2010.

[3] Zelenak M., M., Schulz, Hernando, J., Speaker diarization of broadcast news in Albayzin 2010 evaluation campaign. EURASIP Journal on Audio, Speech, and Music Processing. December 2012.

[4] Tecnoparla Project. Online: http://www.talp.upc.edu/tecnoparla, , accessed on June 2, 2016.

[5] The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan. Online: http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf, accessed on June 2, 2016.

[6] Lleida, E.; Rodriguez-Fuentes, L.J.; Tejedor, J.; Ortega, A.; Miguel, A.; Bazán, V.; Pérez, C.; de Prada, A.; Penagarikano, M.; Varona, A.; et al. An Overview of the IberSpeech-RTVE 2022 Challenges on Speech Technologies. Appl. Sci. 2023, 13, 8577. https://doi.org/10.3390/app13158577