Albayzin Evaluation: IberSPEECH-RTVE 2020 Multimodal Diarization and Scene Description Challenge

Eduardo Lleida[1], Alfonso Ortega[1], Antonio Miguel[1], Virginia Bazán[2], Carmen Pérez[2], Manuel Gómez[2], and Alberto de Prada[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

MD evaluation plan v1 (March 20, 2020)

Abstract. IberSPEECH-RTVE 2020 Multimodal Diarization and scene description evaluation is a new challenge in the ALBAYZIN evaluation series. The evaluation is supported by the *Spanish Thematic Network on Speech Technology* (RTTH) and *Cátedra RTVE Universidad de Zaragoza* and is organized by ViVoLab, Universidad de Zaragoza.The evaluation will be conducted as part of the Iberspeech 2020[3] conference to be held in Valladolid, Spain, from 18 to 20 November 2020.
The multimodal diarization and scene description evaluation consists of segmenting audiovisual documents according to different speakers and faces, linking those segments which originate from the same speaker and face, and optionally label the scenes in terms of 5 descriptor: environment, place, screen, season, and time.

1 Introduction

The multimodal diarization and scene description evaluation consists of segmenting broadcast audiovisual documents according to a closed set of different speakers and faces, linking those segments which originate from the same speaker and face, and optionally label the scenes in terms of 5 descriptor. For this evaluation, a list of characters to recognize will be given jointly with a set of scene descriptors. The rest of characters on the audiovisual document will be discarted for the evaluation purposes. System outputs must give for each segment who is speaking and who is/are in the image from the list of characters. Optionally, the output can include a set of descriptors for each scene. For each character, a set of face pictures and short audiovisual document will be given.

The goal of this challenge is to continue with the Albayzin evaluations based on multimodal information. In this edition, jointly with the speaker and face diarization, we include a scene description evaluation. We want to evaluate the use

[3] https://iberspeech2020.eca-simm.uva.es/

of audiovisual information for speaker and face diarization, and scene description. We encourage participants to use any audiovisual information for diarization and scene description.

For scene description, the audiovisual document must be diarized in terms of 5 scene descriptors:

– The **environment** descriptor includes two labels, *rural* (countryside scenes) and *urbano* (city scenes).
– The **place** descriptor includes two labels, *interior* (indoor scenes) and *exterior* (outdoor scenes).
– The **screen** descriptor includes only a label, *multipantalla* to label scenes containing multiple screens.
– The **season** descriptor includes two labels, *verano* (warm weather scenes) and *invierno* (cold weather scenes).
– The **time** descriptor includes two labels, *dia* (daylight scenes) and *noche* (nightlight scenes)

2 Database description

2.1 RTVE2018DB

RTVE2018[4] database has been divided into 4 partitions, a *train* one, two development partitions *dev1*, *dev2* and finally a *test* partition.
Detailed information about the RTVE2018 database content can be found in the RTVE2018 database description report `http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf`. Only part of the *dev2* and *test* partitions will be used in this challenge. Partition *dev2* contains a 2 h show annotated for multimodal diarization (face and speaker) and enrollment files (pictures, videos, and audio) needed for speaker and face identification. Also, the *test* partition contains two TV shows with a total of 4 h annotated for multimodal diarization and the corresponding enrollment files.

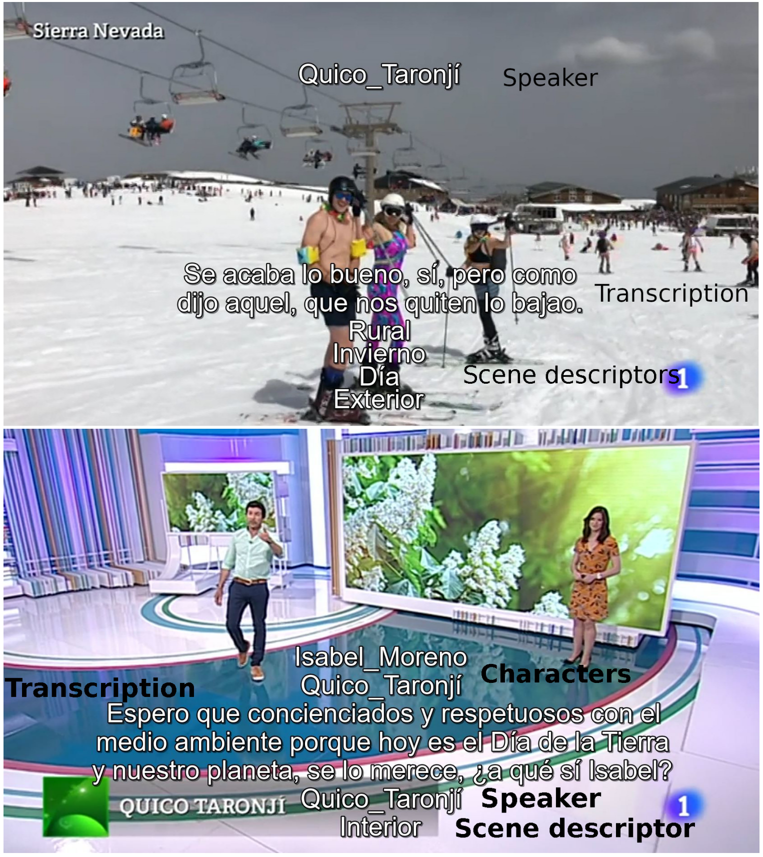
2.2 RTVE2020DB

The RTVE2020 database is a collection of TV shows that belong to diverse genres and broadcast by the public Spanish Television (RTVE) from 2018 to 2019. A total of 33 h and 21 min of audio and video have been labeled in terms of speaker and character turns, and scene descriptors.

More than 175 characters have been labeled and their corresponding enrollment files (pictures and videos with audio) needed for speaker and face identification are provided. Figure 2.1 shows two examples of scenes with the associated transcription, speaker and character names, and the scene descriptors. The enrollment material consists of 10 pictures and a 20 second video with the corresponding audio of each known character.

[4] `http://catedrartve.unizar.es/reto2018.html`

Fig. 1. Examples of the metadata included on the labeled material: transcription, speaker and character names, and scene descriptors


Sierra Nevada
Quico_Taronjí
Speaker
Se acaba lo bueno, sí, pero como dijo aquel, que nos quiten lo bajao.
Transcription
Rural
Invierno
Día
Exterior
Scene descriptors
1
Isabel_Moreno
Quico_Taronjí
Characters
Transcription
Espero que concienciados y respetuosos con el medio ambiente porque hoy es el Día de la Tierra y nuestro planeta, se lo merece, ¿a qué sí Isabel?
QUICO TARONJÍ
Quico_Taronjí
Speaker
Interior
Scene descriptor
1

2.3 Development data.

For development, the dev2 partition of the RTVE2018 database contains a two-hour show “La noche en 24H” labeled with speaker and face timestamps. Enrollment files for the main characters are also provided. Enrollment files consist of pictures and short videos with the character speaking. Additionally, the dev2 partition contains around 14 h of speaker diarization timestamps. Also, the test partition of RTVE2018 database contains three television programs labeled with speaker and face timestamps, one from “La Mañana” and two from “La Tarde en 24H Tertulia”, which totaled four hours. For enrollment, photos (10) and video (20 s) of the 39 characters to be labeled are provided.

Additionally, the RTVE2020 database contains a development partition with around 4 hours labeled with speaker, face and scene descriptors timestamps (see table 1).

No restrictions are placed on the use of any data outside the RTVE2018 and RTVE2020.

Table 1. RTVE2020 development partition for diarization tasks with shows and duration.

Show	Duration
Aquí la tierra	02:56:38
Bajo la red	00:59:02
	03:55:31

2.4 Evaluation data.

The evaluation data will contain a set of TV shows covering a variety of scenarios from RTVE2020 database. The *test* partition contains around 29 h of audiovisual documents labeled in terms of speaker, face and scene descriptors timestamps. The scene descriptors are used to label the video in terms of the scene content. We have defined 5 scene descriptor objects which are: environment, place, screen, season, and time.

- The **environment** descriptor includes two labels, *rural* (countryside scenes) and *urbano* (city scenes).
- The **place** descriptor includes two labels, *interior* (indoor scenes) and *exterior* (outdoor scenes).
- The **screen** descriptor includes only a label, *multipantalla* to label scenes containing multiple screens.
- The **season** descriptor includes two labels, *verano* (warm weather scenes) and *invierno* (cold weather scenes).
- The **time** descriptor includes two labels, *dia* (daylight scenes) and *noche* (nightlight scenes)

More than 175 characters have been labeled and their corresponding enrollment files (pictures and videos with audio) needed for speaker and face identification are provided. Figure 2.1 shows two examples of scenes with the associated transcription, speaker and character names, and the scene descriptors. The enrollment material consists of 10 pictures and a 20 second video with the corresponding audio of each known character.

The detailed information about the evaluation data will be released by June 1st coinciding with the beginning of the evaluation task.

3 Performance Scoring

3.1 Speaker and face diarization

The multimodal diarization performance scoring will evaluate the accuracy of indexing a TV show in terms of the people speaking and present in the image.

To measure the performance of the proposed systems, the Diarization Error Rate (DER) will be computed as the fraction of speaker or face time that is not correctly attributed to that specific character. This score will be computed over the entire file to be processed; including regions where more than one character 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 and face 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 or faces that are present in segment n, $N_{sys}(n)$ is the number of system speakers or faces that are present in segment n and $N_{Correct}(n)$ is the number of reference speakers or faces in segment n correctly assigned by the diarization system.

$$DER = \frac{\sum\limits_{n \in \Omega} E(n)}{\sum\limits_{n \in \Omega} \left(T(n) N_{ref}(n) \right)} \tag{2}$$

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

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

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 s, 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/face begins or ends.

The primary metric to rank systems will be the average of the face and speaker diarization errors

$$DER_{total} = 0.5DER_{spk} + 0.5DER_{face} \tag{3}$$

3.2 Scene description

The performance scoring will evaluate the accuracy of indexing a TV show in terms of the scene descriptors. To measure the performance of the proposed systems, the Diarization Error Rate (DER) will be computed as the fraction that each of the 5 scene descriptors are not correctly attributed. This score will be computed for each descriptor over the entire file to be processed.

The 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 every scene descriptor change points found in both the reference and the hypothesis, and the diarization error time for each segment n is defined as equation 1 and 2. As for the speaker and face diarization, the description diarization error time includes the time that is assigned to the wrong descriptor, missed descriptor time and false alarm descriptor time.

A forgiveness collar of 0.25 s, 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 descriptor begins or ends.

Systems will be ranked in terms of the individual scene descriptor DER and the average of the 5 scene descriptor diarization errors

$$\begin{aligned} DER_{total} = 0.25DER_{environment} + 02.5DER_{place} + \\ 0.25DER_{screen} + 0.25DER_{season} + 0.25DER_{time} \end{aligned} \tag{4}$$

3.3 Segmentation Scoring Tool and System Output Files

The tool used for evaluating the segmentation system is the one developed for the RT Diarization evaluations by NIST “md-eval-v22.pl”, available in the *scoring* folder of the RTVE2020 database distribution.

The format’s definition for the submission of the Multimodal Diarization and Scene Description 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 multimodal diarization system output and reference files.

The RTTM files are space-separated text files that contains meta-data ”Objects” that annotate elements of the recording. Each line represents the annotation of 1 instance of an object.

OBJECT File Channel Beg_Time Dur <NA> <NA> Object_Label <NA> <NA>
Where:

- **OBJECT**: A tag indicating that the segments contains information about the beginning, duration, identity, etc. of a segment that belongs to a certain OBJECT.
- **file**: It is the name of the considered file.
- **tbeg**: The beginning time of the segment, in seconds, measured from the start time of the file.
- **tdur**: It indicates the duration of the segment, in seconds.
- **Object_Label**: It refers to the label asigned to the OBJECT 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 '.'.

Object types can be used or not used depending on the particular evaluation. Table 2 shows the RTTM field names and values used in the RTVE2018 and RTVE2020 databases. A more detailed description of the format can be found in Appendix C of the 2015 KeyWord Search Evaluation Plan[5]. For the sake of clarity new objects have been defined to annotate the face appearances and scene descriptors.

The Multimodal Diarizaton evaluation will use a modified version of md-eval version 22 software to accept the new OBJECTS. We will use the best match between hypothesis and reference labels. The command line will be:
md-eval-v22.pl -c 0.25 -b -r <OBJECT-REFERENCE>.rttm -s <OBJECT-SYSTEM>.rttm

4 General Evaluation Conditions

The organizers encourage the participation of all researchers interested in multimodal diarization and scene description. All teams willing to participate in this evaluation must registered through the challenge web page
`http://catedrartve.unizar.es/albayzin2020.html`
before June 1st, 2020.
In case of any difficulty, you can send an e-mail to lleida@unizar.es

4.1 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

[5] `https://www.nist.gov/sites/default/files/documents/itl/iad/mig/KWS15-evalplan-v05.pdf`

Table 2. RTTM files names used

Field 1	2	3	4	5	6	7	8	9	10
SPKR-INFO	file	1	<NA>	<NA>	<NA>	unknown	speaker_label	<NA>	<NA>
SPEAKER	file	1	tbeg	tdur	<NA>	<NA>	speaker_label	<NA>	<NA>
FACE-INFO	file	1	<NA>	<NA>	<NA>	unknown	face_label	<NA>	<NA>
FACE	file	1	tbeg	tdur	<NA>	<NA>	face_label	<NA>	<NA>
ENVIRON-INFO	file	1	<NA>	<NA>	<NA>	unknown	rural, urban	<NA>	<NA>
ENVIRON	file	1	tbeg	tdur	<NA>	<NA>	environ_label	<NA>	<NA>
PLACE-INFO	file	1	<NA>	<NA>	<NA>	unknown	interior, exterior	<NA>	<NA>
PLACE	file	1	tbeg	tdur	<NA>	<NA>	place_label	<NA>	<NA>
SCREEN-INFO	file	1	<NA>	<NA>	<NA>	unknown	multipantalla	<NA>	<NA>
SCREEN	file	1	tbeg	tdur	<NA>	<NA>	screen_label	<NA>	<NA>
SEASON-INFO	file	1	<NA>	<NA>	<NA>	unknown	inverno, verano	<NA>	<NA>
SEASON	file	1	tbeg	tdur	<NA>	<NA>	season_label	<NA>	<NA>
TIME-INFO	file	1	<NA>	<NA>	<NA>	unknown	dia, noche	<NA>	<NA>
TIME	file	1	tbeg	tdur	<NA>	<NA>	time_label	<NA>	<NA>

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.

4.2 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 or watching to the evaluation data, or any other human interaction with the evaluation data, is not allowed before all results have been submitted. Any publicly available data can be used for training together with the data provided by the organization team. In case of using additional material, the participant will notify it and provide the references of this material.

4.3 Results 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 and modality. The file output file must be identified by the following code:

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

- <**SITE**>: Refers to the 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.
- <**OBJECT**>: Any of the 8 OBJECTS defined in table 2 (SPEAKER, FACE, ENVIRON, PLACE, SEASON, TIME, SCREEN).

Each participant team must upload the zip files through the challenge web page
`http://catedrartve.unizar.es/albayzin2020.html` In case of uploading problems send an e-mail with the corresponding ZIP result files to

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

4.4 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 2020 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, including test results and a post-analysis, can be submitted to the IberSpeech Conference as a regular paper. Please, take advise of the deadlines in the IberSpeech 2020 web page
`https://iberspeech2020.eca-simm.uva.es/`

4.5 Schedule

- March 23th, 2020: Registration opens and release of the training data.
- September 7th, 2020: Registration deadline. Release of the evaluation data.
- October 9th, 2020: Deadline for submission of results and system descriptions.
- October 30th, 2020: Results distributed to the participants.
- December 23th, 2020: Paper submission deadline
- March, 2021: Iberspeech 2020 conference.

5 Acknowledgments

The organizing team would like to thank Corporación Radiotelevisión Española and Cátedra RTVE de la Universidad de Zaragoza for their efford providing the data for the 2020 evaluation. Thanks also to the organizing committee of Iberspeech 2020 for their help and support.