RTVE 2018, 2020 and 2022 Database Description

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

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

RTVE 2018, 2020 & 2022 Databases v1, April 11, 2022.

Abstract. This document presents the RTVE databases. The *Corporación Radiotelevisión Española*[3] and *Cátedra RTVE de la Universidad de Zaragoza* has released a set of databases with audiovisual and textual documents suitable to be used on the Albayzín Evaluation series supported by the *Spanish Thematic Network on Speech Technology* (RTTH)[4]. The database comprises different releases since 2018 and used in the Albayzín Evaluation in 2018 and 2020. This document introduce a new 2022 release for the next 2022 Albayzín Evaluation. The RTVE database contains programs broadcasted by Radiotelevisión Española from the 60's up to 2022. The programs cover a great variety of scenarios from studio to live broadcast, from read speech to spontaneous speech, different Spanish accents, including Latin-American accents and a great variety of contents including fiction series and RTVE historical archive.

1 Introduction

In 2017, the *Corporación Radiotelevisión Española* (RTVE) and the *Universidad de Zaragoza* (UZ) signed an agreement to develop the *"Cátedra RTVE de la Universidad de Zaragoza"* with the purpose of boosting the technologies associated with the generation of audiovisual metadata.

One of the objectives of the Cátedra is to launch a set of technological challenges and provide the necessary data to test the technologies. For this purpose, RTVE released, in 2018, about 586 hours of audio and associated subtitles extracted from different RTVE programs and the whole subtitles broadcasted by the RTVE 24H channel in year 2017. This dataset was used in the IberSpeech-RTVE 2018 challenges. In 2020, a new set of challenges were organized under the umbrella of the IberSpeech 2020 [5]. For this purpose, RTVE released a new dataset of 55 hours of audio and videos with the associated transcriptions. Half of the dataset was labeled in terms of speakers, characters and a set of scene descriptions. This year 2022, a new set of challenges are organized under the umbrella of the IberSpeech 2022 [6]. This year the challenges are restricted to speech technologies, covering Speech to Text, Speaker Diarization and Identity Assignment, Search on Speech and Text and Speech Alignment. RTVE will release a new dateset, RTVE2022DB, of 50 hours of audio with the associated transcriptions, 20 hours with speaker assignment and 10 hours of audio and associated subtitles produced by respeaking.

[3] http://www.rtve.es

[4] http://www.rthabla.es

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

[6] http://iberspeech2022.ugr.es

All databases, RTVE2018DB, RTVE2020DB and RTVE2022DB are available to the 2022 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 [7]

2 Database content

2.1 RTVE2018DB

The RTVE2018 database is a collection of TV shows that belong to diverse genres and broadcast by the public Spanish Television (RTVE) from 2015 to 2018.

The database is composed of 569 h and 22 min of audio. About 460 h are provided with the subtitles, and about 109 h have been human transcribed. We would like to highlight that in most of the cases, subtitles do not contain verbatim transcriptions of the audio since most of them were generated by a re-speaking procedure (The re-speaker re-utters everything that is being said to a speech-to-text transcription system. Most of the time, the re-speaker summarizes what is being said.).

The corpus is divided into four partitions, a training one, two development partitions (dev1 and dev2) and finally, a test partition. Additionally, the corpus includes a set of text files extracted from all the subtitles broadcast by the RTVE 24H Channel during 2017.

The training partition consists of all the audio files without human revised transcriptions, which means that only subtitles are available. The training partition can be used for any evaluation task. For development, two partitions are defined. Partition dev1 contains about 53 h of audio and their corresponding human revised transcriptions. The dev1 partition can be used for either development or training of the speech-to-text systems. Partition dev2 contains about 15 h of audio, human revised transcriptions, speaker changes, and their corresponding speaker labels. Additionally, 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. Table 1 shows detailed information about the shows included in the development partitions.

The RTVE2018 database includes a test partition with all the files needed to evaluate systems for speech-to-text and speaker and multimodal diarization. Table 2 presents the content of the test partition. The test set covers diverse genres from broadcast news, live magazines, quiz shows, to documentary series with a diversity of acoustic scenarios. Additionally, the test partition contains the enrollment files for the multimodal diarization challenge. It consists of 10 pictures and a 20 second video of the 39 characters to be identified.

Table 3 presents the titles, duration, and content of the shows included in the RTVE2018 database.

Note *that in the current version of the database, the transcriptions aren't synchronized with the audio so the time information of the stm files associated to the audio files are based on dummy segments, only diarization files, rttm, in dev2 and test contain exact time marks for speaker turns.*

More info about the use of the RTVE2018 database in the IberSpeech 2018 challenge can be found in [1].

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. Table 4 presents the titles and duration of the shows included in the RTVE2020 database.

[7] http://catedrartve.unizar.es/rtvedatabase.html

The database is composed of 55 h and 40 min of audio. The whole database has been human transcribed and it was used as *test* partition for the 2020 speech to text challenge.

Additionally, 33 h and 21 min of audio and video have been labeled in terms of speaker and character turns, and scene descriptors. 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)

161 characters have been labeled and their corresponding enrollment files (pictures and videos with audio) needed for speaker and face identification are provided. Figure 2.2 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. Around 4 hours of the database are given as development set (see table 6) and the rest, 29 hours, were released as test (see table 5).

All the partitions, *train*, *dev1*, *dev2* and *test* of the RTVE2018DB database and *dev* and *test* of the RTVE2020DB can be used as training or development for the new 2022 challenge.

More info about the use of the RTVE2018DB and RTVE2020DB in the IberSpeech 2020 challenge can be found in https://catedrartve.unizar.es/albayzin2020results.html

Table 1. RTVE2018DB development (dev) partitions with shows and duration. S2T, speech-to-text; SD, speaker diarization; MD, multimodal diarization.

dev1	**Hours**	**Track**	**dev2**	**Hours**	**Track**
20H	9:13:13	S2T			
Asuntos Públicos	8:11:00	S2T			
Comando Actualidad	7:53:13	S2T			
La Mañana	1:30:00	S2T			
			Millennium	7:42:44	SD, S2T
La noche en 24H	25:44:25	S2T	La noche en 24H	7:26:41	SD, S2T, MD
	52:31:51			15:09:25	

Table 2. RTVE2018DB test dataset partition with shows and duration.

Show	**S2T**	**SD**	**MD**
Al filo de lo Imposible	4:10:03		
Arranca en Verde	1:00:30		
Dicho y Hecho.	1:48:00		
España en Comunidad	8:09:32	8:09:32	
La Mañana	8:05:00	1:36:31	1:36:31
La Tarde en 24H (Tertulia)	8:52:20	8:52:20	2:28:14
Latinoamérica en 24H	4:06:57	4:06:57	
Saber y Ganar	2:54:53		
	39:07:15	22:45:20	4:04:45

Table 3. Information about the shows included in the **RTVE2018DB** database.

Show	**Duration**	**Show Content**
20H	41:35:50	News of the day.
Agrosfera	37:34:32	Agrosfera wants to bring the news of the countryside and the sea to farmers, ranchers, fishermen, and rural inhabitants. The program also aims to bring this rural world closer to those who do not inhabit it, but who do enjoy it.
Al filo de lo Imposible	11:09:57	This show broadcasts documentaries about mountaineering, climbing, and other outdoor risky sports. It is a documentary series in which emotion, adventure, sports, and risk predominate.
Arranca en Verde	05:38:05	Contest dedicated to road safety. In it, viewers are presented with questions related to road safety in order to disseminate in a pleasant way the rules of the road and thus raise awareness about civic driving and respect for the environment.
Asuntos Públicos	69:38:00	All the analysis of the news of the day and the live broadcast of the most outstanding information events.
Comando Actualidad	17:03:41	A show that presents a current topic through the choral gaze of several street reporters. Four journalists who travel to the place where the news occurs show them as they are and bring their personal perspective to the subject.
Dicho y Hecho	10:06:00	Game show in which a group of 6 comedians and celebrities compete against each other through hilarious challenges.
España en Comunidad	13:02:59	Show that offers in-depth reports and current information about the different Spanish autonomous communities. It is made by the territorial and production centers of RTVE.
La Mañana	227:47:00	Live show, with a varied offer of content for the whole family and with the clear vocation of public service.
La Tarde en 24H Economia	04:10:54	Program about the economy.
La Tarde en 24H Tertulia	26:42:00	Talk show of political and economic news (4 to 5 people).
La Tarde en 24H Entrevista	04:54:03	In-depth interviews with personalities from different fields.
La Tarde en 24H el Tiempo	02:20:12	Weather information of Spain, Europe, and America.
Latinoamérica en 24H	16:19:00	Analysis and information show focused on Ibero-America, in collaboration with the information services of the international area and the network of correspondents of RTVE.
Millennium	19:08:35	Debate show of ideas that pretends to be useful to the spectators of today, accompanying them in the analysis of everyday events.
Saber y Ganar	29:00:10	Daily contest presented that aims to disseminate culture in an entertaining way. Three contestants demonstrate their knowledge and mental agility, through a set of general questions.

Table 4. Information about the shows included in the **RTVE2020DB** database.

Show	**Duration**	**Show Content**
Ese programa del que usted me habla	01:58:36	A TV program that reviews daily political, cultural, social and sports news from the perspective of comedy.
Los desayunos de TVE	10:58:34	The daily news, politics, interviews and debate program
Neverfilms	00:11:41	A webseries that parody humorously trailers of series and movies well known to the public.
Si fueras tu	00:51:14	Interactive series that tells the story of Alba, a 17-year-old girl who arrives in a new urbanization and discovers that she has a mysterious relationship with a girl who disappeared six months earlier.
Bajo la red	00:59:01	It is a youth fiction series whose plot is about a chain of favors on the internet
Comando Actualidad	4:01:31	A show that presents a current topic through the choral gaze of several street reporters. Four journalists who travel to the place where the news occurs show them as they are and bring their personal perspective to the subject.
Boca norte	01:00:46	A story of young people who dance to the rhythm of trap and set in Barcelona
Wake-up	00:57:28	A story that combines science fiction, a post-apocalyptic Madrid and lots of action inspired aesthetically and narratively in video games.
Versión española	02:29:12	Program dedicated to the promotion of Spanish and Latin American cinema.
Aquí la tierra	10:26:02	A magazine that deals with the influence of climatology and meteorology both personally and globally.
Mercado central	08:39:47	A Spanish soap opera set in a today's Madrid market.
Vaya crack	05:06:00	A contest where contestants take multiple quiza designed to test their abilities in one or more of the following intelligence categories: musical, physical, social, logical, visual, and linguistic.
Cómo nos reímos	02:51:42	A program dedicated to the great comedians and their work on RTVE programs.
Imprescindibles	3:12:31	A documentary series on the most outstanding figures of Spanish culture in the 20th century. The audio corresponeds to raw material and contain to microphone channels, the interviewer and interviewed.
Millennium	1:56:11	Debate show of ideas that pretends to be useful to the spectators of today, accompanying them in the analysis of everyday events.
Total duration	**55:40:16**	

Table 5. RTVE2020DB test partition for diarization tasks with shows and duration.

Show	**Duration**
Aquí la tierra	02:56:38
Los desayunos de TVE	10:58:34
Neverfilms	00:11:42
Si fueras tu	00:51:14
Comando Actualidad	04:01:31
Boca norte	01:00:46
Wake-up	00:57:28
Aqui la tierra	07:19:24
TOTAL	29:25:25

Table 6. RTVE2020DB development partition for diarization tasks with shows and duration.

Show	**Duration**
Ese programa del que usted me habla	01:58:36
Bajo la red	00:59:02
TOTAL	03:55:31

2.3 RTVE2022DB

The RTVE2022DB database is a collection of a diverse audio material recorded from the 60's to the present. It covers from historical recordings, popular broadcated TV shows and fictional shows. The database contains three partitions: a training partition with audio and the corresponding subtitles aligned for training ASR systems, a development partition with audio and the corresponding broadcast subtitles and the reference subtitles for the Text and Speech Alignment evaluation, and finally, a test partition with the audio for all the challenges. The training partition has been prepared by Vicomtech[8] and it contains 260 TV programs of 9 different shows broadcast by RTVE (see Table 7). The training partition is made up of 168 hours of audio transcribed and automatically aligned at the sentence level, that is, including their start and end timestamps.

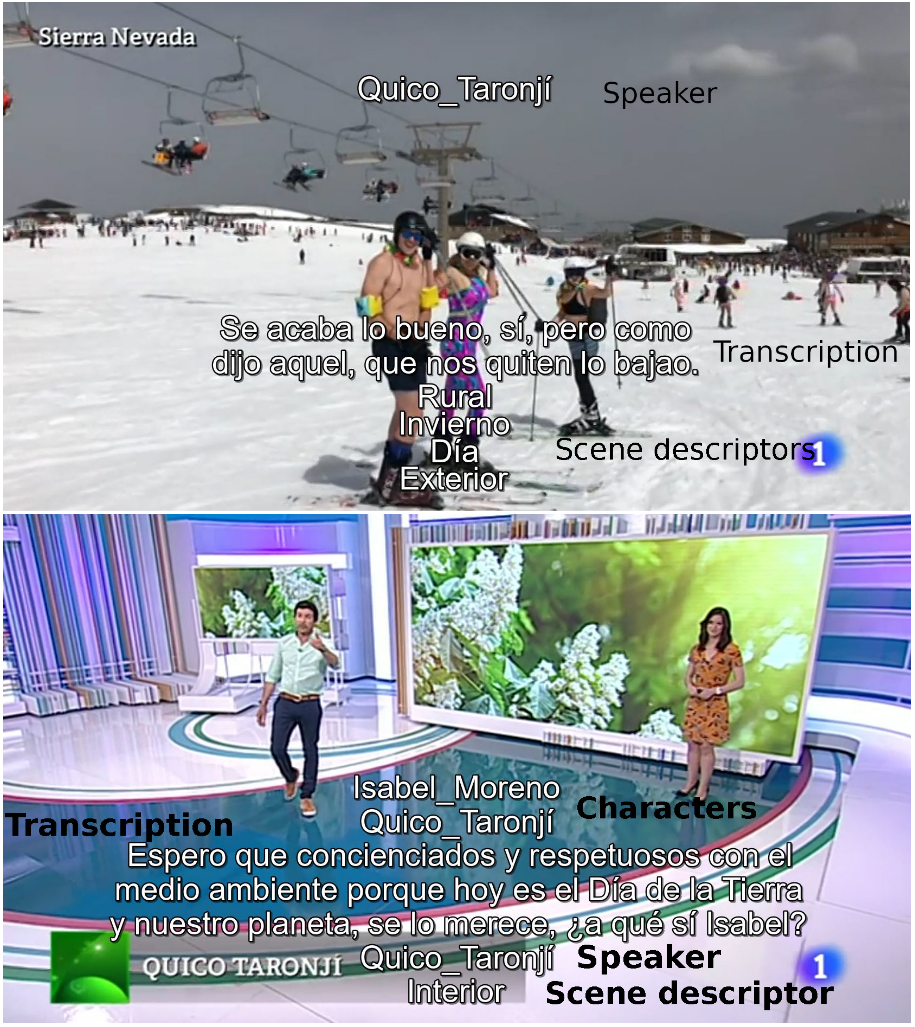
The test partition will made up of 55 hours of diverse audio material. The test partition will be human transcribed and it will be used as ***test*** partition for the 2022 Speech to Text and Search on Speech challenges.

Additionally, around 25 hours of audio has been labeled in terms of speaker turns and assigned an identity to the speakers. We expect to have more than 100 different speakers for identity assignment.

The RTVE2022DB contains new material for the Text and Speech Alignment of subtitles generated by respeaking. As it is a new challenge, we will release a development partition with more than 2 hours of audio and subtitles of 2 TV shows: Aquí la Tierra y Agroesfera. The test material will be made up of the same development TV shows and an additional new one with a total of 10 hours.

[8] https://www.vicomtech.org/

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

2.4 Database structure

The structure of the **RTVE2018DB** database is as follows:

- ***RTVE2018/train*** - a folder with the *train* dataset.
- ***RTVE2018/train/audio*** - a folder with the *train* audio files in AAC[9] format.
- ***RTVE2018/train/srt*** - a folder with the subtitles associated with the training audio files in srt[10] format.
- ***RTVE2018/dev1*** - a folder with the development *dev1* dataset.
- ***RTVE2018/dev1/audio*** - a folder with the development *dev1* dataset audio files.

[9] (LC mp4a), 44100 Hz, stereo, variable bitrate.
See section 3.2

[10] https://es.wikipedia.org/wiki/SubRip
See section 3.3

Table 7. Information about the shows included in the training partition of the **RTVE2022DB** database.

Show	**Duration**	**Show Content**
Aquí la Tierra	08:46	A magazine that deals with the influence of climatology and meteorology both personally and globally.
Días de Cine	14:45	The program looks at the most outstanding premieres of the movie billboard.
El Paisano	15:41	We know different towns of the Spanish geography from the hand of the comedian and actor. The artist will spend 48 hours in the smallest towns in Spain and will learn great things and stories from the residents of each town.
Ese Programa del que Usted me Habla	20:09	A TV program that reviews daily political, cultural, social and sports news from the perspective of comedy.
Españoles en el Mundo	28:11	Program in which Spaniards residing outside of Spain show us their place of residence and their environment.
Hacer de Comer	10:11	Daily cooking program presented by Dani García. Throughout the week, Dani García will teach us how to cook traditional and delicious recipes that we can all do at home.
¿Juegas o Qué?	35:53	Nine comedians will convince pedestrian to participate in their fun games. The hosts will surprise anonymous people who can earn money with their knowledge of general culture.
La Paisana	07:31	The program, conducted by a comedian (Eva Hache), will discover fun stories of the day-to-day life of the people in small towns in Spain.
Masterchef	139:40	Masterchef is a space for self-improvement and effort where the goal of the participants is to achieve their dream: dedicate themselves professionally to the kitchen.

- ***RTVE2018/dev1/trn*** - a folder with the development *dev1* dataset human revised word transcriptions in trn[11] format.
- ***RTVE2018/dev1/stm*** - a folder with the development *dev1* dataset stm[12] reference files for ASR scoring.
- ***RTVE2018/dev2*** - a folder with the development *dev2* dataset.
- ***RTVE2018/dev2/audio*** - a folder with the development *dev2* dataset audio files in AAC format.
- ***RTVE2018/dev2/trn*** - a folder with the development *dev2* dataset human revised word transcriptions in trn format.
- ***RTVE2018/dev2/stm*** - a folder with the development *dev2* dataset stm reference files for ASR scoring.
- ***RTVE2018/dev2/rttm*** - a folder with the development *dev2* dataset reference speaker and face diarization files in rttm[13] format.

[11] each line contains speaker identity (#_speaker) and the word transcriptions
See section 3.4

[12] Reference file format used by sclite NIST scoring tool
See section 3.5

[13] A modified version of the NIST format to include the type object FACE.
See section 3.6

- ***RTVE2018/dev2/video*** - a folder with the development *dev2* dataset audiovisual files in mp4[14] format.
- ***RTVE2018/dev2/enrollment*** - a folder with the development *dev2* dataset enrollment files for person identification in the Speaker and Multimodal Diarization tasks.
- ***RTVE2018/dev2/enrollment/<name>*** - a folder with the development *dev2* enrollment files for <*name*> person. For each person to be identified, a set of pictures and mp4 videos with audio are provided as enrollment information.
- ***RTVE2018/test*** - a folder with the *test* dataset.
- ***RTVE2018/test/audio*** - a folder with the *test* dataset audio files in AAC format.
- ***RTVE2018/test/enrollment*** - a folder with the *test* dataset enrollment files for person identification in the Speaker and Multimodal Diarization tasks.
- ***RTVE2018/test/enrollment/<name>*** - a folder with the *test* enrollment files for <*name*> person. For each person to be identified, a set of pictures and mp4 videos with audio are provided as enrollment information.
- ***RTVE2018/test/references*** - a folder with the *test* reference files. The folder contains subfolders with the rttm and stm reference files.
- ***RTVE2018/subtitles*** - a folder with text files extracted from subtitles.
- ***RTVE2018/subtitles/2017*** - a folder with text files extracted from the subtitles broadcast along 2017 at the RTVE 24H channel. Files are plain text using utf-8 charset. Each line is a sentence.
- ***RTVE2018/scoring*** - a folder with the scoring scripts.
- ***RTVE2018/doc*** - a folder with relevant evaluation information: examples output files, evaluation plans, data organization, README file, license agreement, etc.

The structure of the **RTVE2020DB** database is as follows:

- ***RTVE2020/dev*** - a folder with the development *dev* dataset.
- ***RTVE2020/dev/audio*** - a folder with the development *dev* dataset audio files in AAC format.See section 3.2.
- ***RTVE2020/dev/video*** - a folder with the development *dev* dataset audiovisual files in mp4 format. See section 3.1.
- ***RTVE2020/dev/rttm*** - a folder with the development *dev* dataset reference speaker, face and scene descriptors diarization files in rttm format. See section 3.6.
- ***RTVE2020/dev/enrollment*** - a folder with the development *dev* dataset enrollment files for person (speaker and face) identification.
- ***RTVE2020/dev/enrollment/<name>*** - a folder with the development enrollment files for <*name*> person. For each person to be identified, a set of pictures and mp4 videos with audio are provided as enrollment information.
- ***RTVE2020/test*** - a folder with the *test* dataset.
- ***RTVE2020/test/audio*** - a folder with the *test* dataset audio files in AAC format.
- ***RTVE2020/test/audio/S2T*** - a folder with the *test* audio files for the Speech to Text challenge.
- ***RTVE2020/test/audio/SD*** - a folder with the *test* audio files for the Speaker Diarization and Identity Assignment challenge.
- ***RTVE2020/test/enrollment*** - a folder with the *test* dataset enrollment files for person (speaker and face) identification.
- ***RTVE2020/test/enrollment/<name>*** - a folder with the *test* enrollment files for <*name*> person. For each person to be identified, a set of pictures and mp4 videos with audio are provided as enrollment information.
- ***RTVE2020/test/video*** - a folder with the *test* audiovisual files in mp4 format.

[14] https://es.wikipedia.org/wiki/H.264/MPEG-4_AVC

See section 3.1

- ***RTVE2020/test/references*** - a folder with the *test* reference files. The folder contains subfolders with the rttm and stm reference files.
- ***RTVE2020/scoring*** - a folder with the scoring scripts.
- ***RTVE2020/doc*** - a folder with relevant evaluation information: examples output files, evaluation plans, data organization, README file, license agreement, etc.

The structure of the **RTVE2022DB** database is as follows:

- ***RTVE2022/train*** - a folder with the *train* dataset.
- ***RTVE2022/train/audio*** - a folder with the *train* audio files in AAC[15] format.
- ***RTVE2022/train/stm*** - a folder with the subtitles aligned associated with the training audio files in stm format.
- ***RTVE2022/dev*** - a folder with the development *dev* dataset.
- ***RTVE2022/dev/audio*** - a folder with the development *dev* dataset audio files in AAC format.See section 3.2.
- ***RTVE2022/dev/audio/TaSA*** - a folder with the *dev* audio files for the Text and Speech Alignment challenge.
- ***RTVE2022/dev/stm*** - a folder with the development *dev* dataset of original subtitles (*subtitles* folder) and reference ones (*ref* folder) in stm format. See section 3.5.
- ***RTVE2022/test*** - a folder with the *test* dataset.
- ***RTVE2022/test/audio*** - a folder with the *test* dataset audio files in AAC format.
- ***RTVE2022/test/audio/S2T*** - a folder with the *test* audio files for the S2T challenge.
- ***RTVE2022/test/audio/SD*** - a folder with the *test* audio files for the Speaker Diarization and Identity Assignement challenge.
- ***RTVE2022/test/audio/SoS*** - a folder with the *test* audio files for the Search on Speech challenge.
- ***RTVE2022/test/audio/TaSA*** - a folder with the *test* audio files for the Text and Speech Alignment challenge.
- ***RTVE2022/test/enrollment*** - a folder with the *test* dataset enrollment audio files for person identification.
- ***RTVE2022/test/enrollment/<name>*** - a folder with the *test* enrollment files for <*name*> person. For each person to be identified, a set of audios are provided as enrollment information.
- ***RTVE2022/test/references*** - a folder with the *test* reference files. The folder contains subfolders with the rttm and stm reference files for each challenge.
- ***RTVE2022/scoring*** - a folder with the scoring scripts.
- ***RTVE2022/doc*** - a folder with relevant evaluation information: examples output files, evaluation plans, data organization, README file, license agreement, etc.

3 Database file formats

RTVE databases contain a set of video, audio and text files. All video and audio files are distributed encoded using the ***mp4*** standard. All the text files are distributed using the ***utf-8*** charset.

3.1 Video files (.mp4)

For multimodal diarization task, development and test video files are provided with the audio track in a mp4 container.

[15] (LC mp4a), 44100 Hz, stereo, variable bitrate.
See section 3.2

The default format is the one used by the on demand Internet channel *”RTVE a la carta”*[16]. The video stream is encoded using the h264 video coding standard with yuv420p pixel format, aspect ratio 1024x576 [SAR 1:1 DAR 16:9], 25 fps and an average bit rate of 1500 kb/s.
The audio stream is encoded using the mpeg Low Complexity (AAC-LC) audio codec with a sampling rate 44100 Hz, stereo and a variable bit rate ranging from 48 to 96 kb/s

3.2 Audio files (.aac)

All the audio files are provided encoded in the AAC format. The stereo audio signal at 44100 Hz sampling rate per channel has been encoded using the mp4-LC profile with a variable bit rate ranging from 48 to 96 kb/s The audio files have been created by extracting the audio stream from the video files without decoding/encoding using the following ffmpeg command:

```
ffmpeg -i <name> -vn -acodec copy ‘basename <name> .mp4‘.aac
```

where <name> is the mp4 video file containing the audio stream to extract.

3.3 Subtitles files (.srt)

The subtitles files are distributed in Subrip format. The Subrip format is a text file with *.srt* extension[17]. The Subrip format consists of four parts, all in plain text:

1. A number indicating which subtitle it is in the sequence.
2. The time that the subtitle should appear on the screen, and then disappear.
3. The subtitle itself.
4. A blank line indicating the start of a new subtitle.

Here is an example of a Subrip file:

```
1
00:00:10,000 --> 00:00:13,560
Escuchar el ruido,

2
00:00:13,640 --> 00:00:18,600
hay que escucharlos todos los días.

3
00:00:22,560 --> 00:00:25,320
-La satisfacción de una isla
que está desierta

4
00:00:25,360 --> 00:00:30,360
y va a una expedición y puedes
hacerla con los medios que tenemos.
```

Subrip files are easily manipulated using the pysrt[18] library in Python.

[16] http://www.rtve.es/alacarta/

[17] https://matroska.org/technical/specs/subtitles/srt.html

[18] https://github.com/byroot/pysrt

3.4 Reference Trancription files (.trn)

The human-revised word transcriptions are given in text files. The transcription files use the ***.trn*** extension. A TRN format consists of text lines with a speaking turn structure. Each line is a turn beginning with the speaker id (#_<ID>) and the word transcriptions.
Here is an example:

(#_0) Abuelo.
(#_1000) (Gritan todos) ¡Abuelo!
(#_1) ¿De dónde vienen ustedes?
(#_2) De Galicia.
(#_1) Vienen bien lejos, entonces, aquí a conocer El Torcal
(#_3) A ver lo más bonito que tienen aquí.
(#_4) ¿Están ustedes esperando para subir al castañar?
(#_1000) (Gritan niños) Sí.
(#_4) ¿Y nos dicen que llevan cuánto tiempo?
(#_5) Hora y media.
(#_6) ¿Y toda esta gente a qué viene?
(#_7) Vienen a ver la berrea.
(#_1000) (Bramido)
(#_4) Es ahora o nunca. Yo que usted me perdería en ellos.
(#_8) Los meses fuertes son los meses de primavera y de otoño.
(#_6) ¿Qué significa, entonces, para ti este este paraje?
(#_9) Es uno de los sitios más bonitos que he visto.
(#_10) Es una maravilla, con los colores ocres y...

The speaker ID (#_1000) is used as special speaker id mark for relevant non-speech turns as music, laughter, shouting and so on. The non-speech audio is written in parentheses, as *(Gritan todos)*[19].

3.5 ASR reference files (.stm)

The STM format describes the segment time marked files consisting of a concatenation of text segment records from a waveform file[20]. Each record is separated by a newline and contains: the waveform's filename and channel identifier [A|B], the talkers ID, begin and end times (in seconds), optional subset label and the text for the segment. Here is an example of stm file:

20H 1 Presentador1 2079.102 2086.618 <,,> El premio se les concedió por sus descubrimientos sobre los mecanismos moleculares que controlan los ritmos cardiacos
20H 1 Presentador2 2086.642 2092.578 <,,> En la información que van a ver a continuación van a intentar explicar qué es exactamente eso .
20H 1 Voz_off8 2093.900 2101.040 <,,> Los ritmos circadianos podrían traducirse popularmente como los mecanismos de nuestro reloj biológico interno

3.6 Diarization files (.rttm)

For speaker and multimodal diariation task, the dataset contains Rich Transcription Time Marked (RTTM) files with the ground-truth. 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 types can be used or not used depending on the particular evaluation. Table 8 shows the RTTM field names and values used in the RTVE2018

[19] all screaming
[20] http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/infmts.htm

and RTVE2020 databases. A more detailed description of the format can be found in Appendix C of the 2015 KeyWord Search Evaluation Plan[21]. For the sake of clarity new objects have been defined to annotate the face appearances and scene descriptors.

Table 8. 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>

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 '.'.

4 License

The RTVE data is available to the IberSPEECH-RTVE 2022 Challenge evaluation participants 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 and send a scanned copy attached to the email. A digital signature is also valid. A copy signed by RTVE representative will be returned following the instructions given in the web page. RTVE authorize the use of the contents released for the call IBERSPEECH-RTVE Challenge 2022, for its use in research works, to all those participants.

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

The authorization will be valid for three years from the date of the public communication of the results of the Challenge 2022. After this period, if necessary, an extension may be requested for the same use.

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. Applied Sciences, Vol 9, Num 24, 2019, doi=10.3390/app9245412