RTVE2018 Database Description

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,ivinalsb,amiguel,lleida}@unizar.es
http://www.vivolab.es
[2] Corporación Radiotelevisión Española, Spain
http:www.rtve.es

RTVE2018 Database v1, June 18, 2018.

Abstract. This document presents the RTVE2018 database. The *Corporación Radiotelevisión Española*[3] and *Cátedra RTVE de la Universidad de Zaragoza* has released a database with audiovisual and textual documents suitable to be used on the ALBAYZIN evaluation series supported by the *Spanish Thematic Network on Speech Technology* (RTTH)[4]. The database comprises different programs broadcast by Radiotelevisión Española from 2015 to 2018. 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.

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 releases about 586 hours of audio and associated subtitles extracted from different RTVE programs and the whole subtitles broadcast by the RTVE 24H channel in year 2017. Additionally, and thanks to the RTTH, more than 100 hours of audio have been transcribed and human revised. The data is available to the evaluation participants only 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/reto2018.html

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

2 Database content

RTVE2018 database is a collection of whole TV shows drawn from diverse genres and broadcast by the public Spanish National Television (RTVE) from 2015 to 2018. Table 1 presents the titles, duration and content of the shows included on the RTVE2018 database.

There are a total of 569 hours and 22 minutes of audio. About 460 hours are provided with the subtitles and about 109 hours have been human-revised transcribed. Be aware that in most of the cases, subtitles could not contain an accurate word transcription as most of them have been generated by a respeaking procedure.[5]

The database has been divided in 4 partitions, a *train* one, two development partitions *dev1*, *dev2* and finally a *test* partition. Additionally, the database includes a set of text files extracted from all the subtitles broadcast by the RTVE 24H Channel during 2017.

The train partition consists of all the audio files without human-revised transcriptions, which means that only subtitles are available. The train partition can be used for any evaluation task.

For development, two partitions have been defined. Partition *dev1* contains about 53 hours of audios and their corresponding human-revised transcriptions. *Dev1* partition can be used for either development or training speech to text systems. Partition *dev2* contains about 15 hours of audios, human-revised transcriptions and diarization files. Additionally, *dev2* contains a 2 hours show annotated for multimodal diarization (face and speaker) and enrollment files (pictures, videos and audios) needed for speaker and face identification.

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 contain exact time marks for speaker turns.*

Table 2 shows more detailed information about the shows included on the development partitions.

RTVE2018 database includes a *test* partition with all the files needed to evaluate systems for speech to text, speaker and multimodal diarization and search on speech. The detailed information about the *partition* will be released with the evaluation data at the middle of September.

2.1 Database structure

The structure of the database is as follows:

– ***RTVE2018/train*** - a folder with the *train* dataset.

[5] The respeaker reutters everything that is being said to a speech to text transcription system. Most of the time the respeaker summarizes what is being said.

Table 1. Information about the shows included in the RTVE2018 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 they do enjoy.
Al filo de lo imposible	11:09:57	This show broadcasts documentaries about mountaineering, climbing and other outdoor risk 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 presented. 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 Magazine, with a varied offer of contents for the whole family and with clear vocation of public service.
La tarde en 24H Economia	04:10:54	Program about economy
La tarde en 24H Tertulia	26:42:00	Talk show of political and economic news. (4/5 people)
La tarde en 24H Entrevista	04:54:03	In-depth interview 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.
La noche en 24H	33:11:06	Talk show with the best analysts to understand what has happened throughout the day. It contains interviews with some of the protagonists of the day.
Total duration	**569:22:04**	

Table 2. Development dataset partition with shows and duration. (S2T: Speech to Text, SoS: Search on Speech)

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	Diarización, S2T, SoS
La noche en 24H	25:44:25	S2T	La noche en 24H	7:26:41	Diarización, S2T, SoS Multimodal
	52:31:51			**15:09:25**	

- ***RTVE2018/train/audio*** - a folder with the *train* audio files in AAC[6] format.
- ***RTVE2018/train/srt*** - a folder with the subtitles associated with the training audio files in srt[7] format.
- ***RTVE2018/dev1*** - a folder with the development *dev1* dataset.
- ***RTVE2018/dev1/audio*** - a folder with the development *dev1* dataset audio files.
- ***RTVE2018/dev1/trn*** - a folder with the development *dev1* dataset human revised word transcriptions in trn[8] format.
- ***RTVE2018/dev1/stm*** - a folder with the development *dev1* dataset stm[9] 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[10] format.
- ***RTVE2018/dev2/video*** - a folder with the development *dev2* dataset audiovisual files in mp4[11] format.

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

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

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

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

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

[11] https://es.wikipedia.org/wiki/H.264/MPEG-4_AVC
See section 3.1

- ***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* dataset 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 *dev2* dataset 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/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.

3 Database file formats

RTVE2018 database contains 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 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.
The default format is the one used by the on demand Internet channel *"RTVE a la carta"*[12]. 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

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

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[13]. 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 das.

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

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

Subrip files are easily manipulated using the pysrt[14] library in Python.

[13] https://matroska.org/technical/specs/subtitles/srt.html
[14] 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)*[15].

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[16]. 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 .

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

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 development *dev2* 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 3 shows the RTTM field names and values used in the RTVE2018 database. A more detailed description of the format can be found in Appendix C of the 2015 KeyWord Search Evaluation Plan[17]. For the sake of clarity an object named FACE has been defined to annotate the face appearances as it is used SPEAKER for speakers turns annotation.

Table 3. 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>

SPEAKER File Channel Beg_Time Dur <NA> <NA> Speaker_Label <NA> <NA>
Where:

- **SPEAKER/FACE**: A tag indicating that the segments contains information about the beginning, duration, identity, etc. of a segment that belongs to a certain speaker/face.
- **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.
- **Speaker/face_Label**: It refers to the label asigned to the speaker/face 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 '.'.

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

4 License

The RTVE data is available to the IberSPEECH-RTVE 2018 Challenge evaluation participants only 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/reto2018.html). Participants must sign the agreement and send a scanned copy attached to the email. A copy signed by RTVE representative will be returned. At the end of the challenge, the licensee shall send a written document to RTVE guaranteeing that all the material has been destroyed, unless he or she has applied for a license extension. RTVE will authorize the use of the contents released for the call IBERSPEECH-RTVE Challenge 2018, for its use in research works, to all those participants who request it. The authorization will be valid for three years from the date of the public communication of the results of the Challenge 2018. After this period, if necessary, an extension may be requested for the same use. Applications must be sent by mail to

RTVE
Dir. FONDO DOCUMENTAL RTVE
Avda. Radiotelevisión, 4
28223 Pozuelo de Alarcón
Madrid
España

The application must indicate user, purpose of research and period of use.