















Article

# An Overview of the IberSpeech-RTVE 2022 Challenges on Speech Technologies

Eduardo Lleida <sup>1,\*</sup> , Luis Javier Rodriguez-Fuentes <sup>2</sup> , Javier Tejedor <sup>3</sup> , Alfonso Ortega <sup>1</sup> , Antonio Miguel <sup>1</sup> , Virginia Bazán <sup>4</sup> , Carmen Pérez <sup>4</sup> , Alberto de Prada <sup>4</sup> , Mikel Penagarikano <sup>2</sup> , Amparo Varona <sup>2</sup> , Germán Bordel <sup>2</sup> , Doroteo Torre-Toledano <sup>5</sup> , Aitor Álvarez <sup>6</sup>  and Haritz Arzelus <sup>6</sup> 

- <sup>1</sup> Vivolab, Aragon Institute for Engineering Research (I3A), University of Zaragoza, 50018 Zaragoza, Spain; ortega@unizar.es (A.O.); amiguel@unizar.es (A.M.)
  - <sup>2</sup> Department of Electricity and Electronics, Faculty of Science and Technology, University of the Basque Country (UPV/EHU), Barrio Sarriena, 48940 Leioa, Spain; luisjavier.rodriguez@ehu.eus (L.J.R.-F.); mikel.penagarikano@ehu.eus (M.P.); amparo.varona@ehu.eus (A.V.); german.bordel@ehu.eus (G.B.)
  - <sup>3</sup> Institute of Technology, Universidad San Pablo-CEU, CEU Universities, Urbanización Montepríncipe, 28668 Boadilla del Monte, Spain; javier.tejedornoguerales@ceu.es
  - <sup>4</sup> Corporación Radiotelevisión Española, 28223 Madrid, Spain; virginia.bazan@rtve.es (V.B.); carmen.perez.cernuda@rtve.es (C.P.); alberto.deprada@rtve.es (A.d.P.)
  - <sup>5</sup> AUDIAS, Electronic and Communication Technology Department, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Av. Francisco Tomás y Valiente, 11, 28049 Madrid, Spain; doroteo.torre@uam.es
  - <sup>6</sup> Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia-San Sebastián, Spain; aalvarez@vicomtech.org (A.Á.); harzelus@vicomtech.org (H.A.)
- \* Correspondence: lleida@unizar.es

**Abstract:** Evaluation campaigns provide a common framework with which the progress of speech technologies can be effectively measured. The aim of this paper is to present a detailed overview of the IberSpeech-RTVE 2022 Challenges, which were organized as part of the IberSpeech 2022 conference under the ongoing series of Albayzin evaluation campaigns. In the 2022 edition, four challenges were launched: (1) speech-to-text transcription; (2) speaker diarization and identity assignment; (3) text and speech alignment; and (4) search on speech. Different databases that cover different domains (e.g., broadcast news, conference talks, parliament sessions) were released for those challenges. The submitted systems also cover a wide range of speech processing methods, which include hidden Markov model-based approaches, end-to-end neural network-based methods, hybrid approaches, etc. This paper describes the databases, the tasks and the performance metrics used in the four challenges. It also provides the most relevant features of the submitted systems and briefly presents and discusses the obtained results. Despite employing state-of-the-art technology, the relatively poor performance attained in some of the challenges reveals that there is still room for improvement. This encourages us to carry on with the Albayzin evaluation campaigns in the coming years.

**Keywords:** IberSpeech Challenge; RTVE2022 database; Albayzin evaluations; speech-to-text transcription; speaker diarization and identity assignment; text and speech alignment; search on speech



**Citation:** 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>

Academic Editor: Douglas O’Shaughnessy

Received: 23 June 2023  
Revised: 19 July 2023  
Accepted: 20 July 2023  
Published: 25 July 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The Albayzin evaluations are a series of technological benchmarks and challenges open to the scientific community within different fields of the broad area of speech technologies. Organized every two years since 2006 and supported by the Spanish Thematic Network on Speech Technologies (Red Temática en Tecnologías del Habla (RTTH)), (<http://www.rthth.es> (accessed on 24 July 2023)), starting from 2018 the evaluations have focused on the broadcast media area. Thanks to Radio Televisión Española, RTVE (<http://www.rtve.es> (accessed on 24 July 2023)), the Spanish Public Broadcast Corporation, and the RTVE Chair at the University of Zaragoza (Cátedra RTVE de la Universidad de Zaragoza: (<http://www.rtve.es> (accessed on 24 July 2023))).

[//catedrartve.unizar.es](https://catedrartve.unizar.es) (accessed on 24 July 2023)), new and more challenging datasets are being released to support the assessment of speech technologies in broadcast media tasks. Speech technologies have been introduced in this area due to their high potential to automate processes as subtitling and caption generation and alignment, or automatic metadata generation for audiovisual content. This requires continuous efforts to evaluate the performance of these technologies and to support new developments. In the last years, deep learning approaches have completely changed the landscape with a great boost in performance in image recognition, natural language processing and speech recognition applications. This has allowed the introduction of speech technologies in the work pipeline of broadcast archives and documentation services.

Different evaluation campaigns for broadcast speech have been carried out from the first one proposed in 1996 [1], most of them using English as the target language [2,3]. In the past few years, speech evaluations organized by the National Institute of Standards and Technology (NIST) have expanded the range of languages of interest (sometimes including low-resourced languages such as Cantonese, Pashto, Tagalog, Swahili, Tamil, to name a few) and speech domains, such as telephone and public safety communication. These evaluations included automatic speech recognition (ASR), spoken term detection and keyword spotting tasks [4–10]. The Albayzin evaluations have always focused on Iberian languages, mainly Spanish and to a lesser extent Catalan and Basque. Spanish is the second language with the highest number of native speakers and the fourth most spoken language in the world (<https://www.ethnologue.com/insights/ethnologue200/> (accessed on 24 July 2023)). The Albayzin evaluations have the purpose of tracking the evolution of speech technologies for Iberian Languages with new and increasingly challenging datasets in each evaluation campaign. In the last editions, the evaluations have focused on Spanish language in the broadcast media sector. In the 2022 edition, four challenges have been proposed:

- Speech to Text Challenge (S2TC), organized by RTVE and Universidad de Zaragoza, which consists of automatically transcribing different types of TV shows.
- Speaker Diarization (SDC) and Identity Assignment Challenge (SDIAC), organized by RTVE and Universidad de Zaragoza, which consists of segmenting broadcast audio documents according to different speakers, linking those segments which originate from the same speaker and identifying a closed set of speakers.
- Text and Speech Alignment Challenge (TaSAC), which consists of two sub-challenges:
  - (Task 1) TaSAC-ST: Alignment of re-speaking generated subtitles, organized by RTVE and Universidad de Zaragoza, which consists of synchronizing the broadcast subtitles created by re-speaking of different TV shows.
  - (Task 2) TaSAC-BP: Alignment and validation of speech signals with partial and inaccurate text transcriptions, organized by the University of the Basque Country (UPV/EHU), which consists of aligning text and audio extracted from a plenary session of the Basque Parliament.
- Search on Speech Challenge (SoSC), organized by Universidad San Pablo-CEU and AuDIaS from Universidad Autónoma de Madrid, which consists of searching in audio content a list of terms/queries.

The main novelty of the IberSpeech-RTVE 2022 Challenges compared to previous evaluations [11,12] relies on the two following features: (1) A text and speech alignment challenge has been proposed for the very first time in the Albayzin evaluation series, including two different sub-tasks for two possible applications; and (2) new and more challenging databases have been released for all the evaluation tasks, covering different domains but focusing on broadcast media content, some of them, such as the RTVE and the Basque Parliament databases, being specifically created for these challenges.

A total of 7 teams from industry and academia registered to participate in the IberSpeech-RTVE 2022 challenges, with 20 different systems submitted in total. Compared to previous evaluations, the number of participants has significantly decreased. In the 2018 edition,

16 teams submitted 85 systems in the different challenges. In the 2020 edition, 12 teams submitted a total of 31 systems.

We present in this paper an overview of the IberSpeech-RTVE 2022 challenges along with the data supplied by the organization to the participants and the performance metrics. The overview includes a detailed description of the systems showcased for evaluation, their corresponding results, and a comprehensive set of conclusions derived from the 2022 evaluation campaign and previous campaigns in 2018 [11] and 2020 [12]. This paper will serve as reference for anyone wanting to use the datasets provided in the evaluation. Up to the date of writing this paper, more than 50 international research groups have asked for access to the RTVE database.

The paper is organized as follows: Section 2 presents the databases used in the different challenges; Section 3 describes the four IberSpeech-RTVE 2022 challenge tasks: speech-to-text transcription, speaker diarization and identity assignment, text and speech alignment, and search on speech, along with with the performance metrics used in each challenge; Section 4 provides a brief description of the submitted systems; Section 5 presents and discusses the results; and finally, a summary of the paper, conclusions and future work are outlined in Section 6.

## 2. IberSpeech-RTVE 2022 Evaluation Databases

### 2.1. RTVE Database

The RTVE database is a collection of TV shows that belong to diverse genres and were broadcast by the public Spanish Television (RTVE). The database has been built incrementally for the 2018, 2020 and 2022 Albayzin challenges resulting in RTVE2018DB, RTVE2020DB and RTVE2022DB datasets. The RTVE database contains nearly 1000 h of broadcast content and subtitles comprising 54 programs of diverse genres and topics, produced and broadcast by RTVE. These programs present various challenges for speech technologies, including the diversity of Spanish accents and slang, overlapping speech, spontaneous speech, acoustic variability, background noise, and specific vocabulary. Table 1 shows a summary of the content of the RTVE database in terms of genres, TV show names, dataset, speech and transcribed hours. The primary aim of the RTVE database is to offer challenging data for the Albayzin evaluations, allowing us to observe the progress of speech technologies when applied to audiovisual content.

For the IberSpeech-RTVE 2022 evaluation, the RTVE2018DB and RTVE2020DB datasets were provided for training and development purposes, while the RTVE2022DB dataset contains additional training, development and test partitions.

The RTVE2018DB dataset comprises complete TV shows spanning a variety of genres that were broadcast on Spain's public National Television from 2015 to 2018. The audio collection amounts to 570 h, of which roughly 460 h come with subtitles, approximately 109 h have undergone human-revised transcription, and 38 h have been labeled with the speaker turns. Additionally, the database features a set of text files that were extracted from all the subtitles broadcast on the RTVE 24H Channel throughout 2017. The RTVE2018 dataset includes partitions with all the files needed to evaluate systems for speech-to-text and speaker and multimodal diarization. A full account of the RTVE2018 dataset content and formats can be found in the dataset description report [13], and its use in the IberSpeech 2018 challenge is described in [11].

The RTVE2020DB dataset is a compilation of complete TV shows from various genres that were broadcast on Spain's public National Television between 2018 and 2019. It contains a total of 56 h of audio, which have been transcribed and reviewed by humans. Furthermore, a subset of 33 h has been categorized according to speaker, face, and scene descriptors. The RTVE2020 dataset includes partitions with all the files needed to evaluate systems for speech-to-text, speaker diarization and identity assignment and multimodal diarization and scene description. Further details about the RTVE2020 dataset content and formats can be found in the dataset description report [14], and information about its use in the IberSpeech 2020 challenge can be found in [12].

Table 1. RTVE database.

Genre	TV Show Name	2018	2020	2022	Total Speech	Transcribed
Cooking show	A pedir de boca			X	3:41:38	3:41:38
Cooking show	Hacer de comer			X	10:11:00	10:11:00
Cooking show	Masterchef			X	139:40:00	139:40:00
Daily magazine	Aquí la tierra		X	X	21:58:46	21:58:46
Daily magazine	Corazón			X	3:00:17	3:00:17
Daytime television	La mañana	X			227:47:00	9:35:00
Daytime televisión	Los desayunos de tve		X		10:58:34	10:58:34
Game show	¿Juegas o qué?			X	35:53:00	35:53:00
Game show	3 × 4			X	2:58:17	2:58:17
Game show	Arranca en Verde	X			5:38:05	1:00:30
Game show	Dicho y Hecho	X			10:06:00	1:48:00
Game show	El cazador			X	3:48:22	3:48:22
Game show	Saber y Ganar	X		X	33:28:38	7:23:21
Game show	Vaya crack		X		5:06:00	5:06:00
Interviews (indoor)	Ateneo			X	1:40:09	1:40:09
Interviews (indoor)	Conversatorios en Casa América			X	1:58:44	1:58:44
Interviews (indoor)	Entrevistas en estudio			X	3:54:57	3:54:57
Interviews (indoor)	Imprescindibles		X		3:12:21	3:12:21
Interviews (street)	Encuestas con ruido ambiente			X	2:08:13	2:08:13
News show	20H	X			41:35:50	9:13:13
News show	Asuntos publicos	X			69:38:00	8:11:00
News show	España en comunidad	X			13:02:59	8:09:32
News show	Informativos UMATIC			X	0:59:49	0:59:49
News show	La tarde en 24H Economía	X			4:10:54	0:00:00
News show	La tarde en 24H El tiempo	X			2:20:12	0:00:00
News show	La tarde en 24H Entrevista	X			4:54:03	0:00:00
News show	La tarde en 24H Tertulia	X			26:42:00	8:52:20
News show	Latinoamerica en 24H	X			16:19:00	4:06:57
News show	Noticias Nacional			X	2:14:32	2:14:32
Outdoor risky sports	Al filo de lo imposible	X			11:09:57	4:10:03
Reality show	Comando actualidad	X	X	X	25:04:41	15:54:13
Reality show	El paisano			X	15:41:00	15:41:00
Reality show	España Directo			X	4:05:57	4:05:57
Reality show	Españoles en el mundo			X	28:11:00	28:11:00
Reality show	La paisana			X	7:31:00	7:31:00
Serial drama	Bajo la red		X		0:59:01	0:59:01
Serial drama	Boca norte		X		1:00:46	1:00:46
Serial drama	Grasa			X	1:29:36	1:29:36

Table 1. Cont.

Genre	TV Show Name	2018	2020	2022	Total Speech	Transcribed
Serial drama	Neverfilms		X		0:11:41	0:11:41
Serial drama	Riders			X	1:15:00	1:15:00
Serial drama	Si fueras tu		X		0:51:14	0:51:14
Serial drama	Wake-up		X		0:57:28	0:57:28
Serial drama	Yreal			X	1:08:46	1:08:46
Sketch comedy	Como nos reíamos		X		2:51:42	2:51:42
Soap opera	Mercado central		X		8:39:47	8:39:47
Talk show	Días de Cine			X	14:45:00	14:45:00
Talk show	Ese programa del que usted me habla		X	X	22:07:36	22:07:36
Talk show	La noche en 24H	X			33:11:06	33:11:06
Talk show	Millennium	X	X		21:04:46	9:38:55
Talk show	Versión española		X		2:29:12	2:29:12
Thematic magazine	Agrosfera	X		X	41:49:52	4:15:20
Thematic magazine	Cerámica Popular Española			X	1:02:35	1:02:35
Thematic magazine	Jara y Sedal			X	2:29:17	2:29:17
Thematic magazine	Toros			X	0:49:57	0:49:57
TOTAL HOURS					960:05:17	497:31:44

The RTVE2022DB dataset is a collection of diverse audio material recorded from the 60 s to the present. It covers historical recordings, popular TV shows and fictional shows. It contains a total of 335 h of audio, which have been transcribed and reviewed by humans. Up to 280 h of speech, corresponding to 260 TV programs of 9 different shows broadcast by RTVE, have been automatically aligned at the sentence level thanks to Vicomtech (<https://www.vicomtech.org/> (accessed on 24 July 2023)). The RTVE2022DB dataset includes partitions with all the files needed to evaluate ASR systems, speaker diarization and identity assignment, text and speech alignment and search on speech. A small development partition is provided for the text and speech alignment challenge.

The audio files have been created by extracting the audio stream from the video files provided by RTVE without decoding/encoding. All the audio files are encoded in the AAC format. Stereo audio signals at 44,100 Hz sampling rate per channel have been encoded using the mp4-LC profile with a variable bit rate ranging from 48 to 96 kb/s.

The RTVE database is freely available subject to the terms of a license agreement with RTVE (<http://catedrartve.unizar.es/rtvedatabase.html> (accessed on 24 July 2023)).

#### RTVE2022DB Test Dataset

The test partition is made up of 21 different TV shows covering 54 h of diverse audio material (See Table 2). It has been human transcribed and, additionally, around 25 h of audio have been labeled in terms of speaker turns and assigned an identity from a closed set of 74 speakers. As enrollment for each speaker, an audio recording of at least 30 s is provided. Table 3 shows the distribution of the dataset among the different challenges.

**Table 2.** Show content and genre of the TV shows included in the 2022 test dataset.

RTVE2022		
Program	Genre	Show Content
3 × 4	Game show	Magazine contest broadcast on TVE from 1986 to 1990, it took place live, except for certain interviews and performances by invited artists, which were recorded in advance.
A pedir de boca	Cooking show	Raw footage of the TV show. The program takes a tour of the history, habitat and processes of making quality food produced in Spain.
Agrosfera	Thematic magazine	Informative program of public service and citizen participation on the news of the primary sector, rural areas and the food industry.
Aquí la Tierra	Daily magazine	A magazine that deals with the influence of climatology and meteorology both personally and globally.
Ateneo	Thematic magazine	Cultural program on art and books broadcast in the 60 s.
Cerámica Popular Española	Thematic magazine	Cultural program on ceramics in rural Spain broadcast in the 80 s.
Comando Actualidad	Reality show	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.
Conversatorios en Casa América	Interviews (indoor)	An interview program with renowned figures that seeks to delve into the richness and diversity of Latin American societies. Guest and journalist will talk in the halls of Casa de América.
Corazón	Daily magazine	Show in which you can find news about the social life of celebrities, fashion, beauty and other current issues.
El cazador	Game show	Four strangers work as a team to answer general knowledge questions. They must defeat the Hunter, a ruthless quiz show genius, to win the prize money.
Encuestas con ruido ambiente	Interviews (street)	Raw footage of interviews in the street.
Entrevistas en estudio	Interviews (indoor)	Raw footage of indoor interviews.
España Directo	Reality show	A life magazine that makes a social chronicle of Spain, getaways, the weather, parties, reports and cooking recipes, as well as lots of entertainment.
Grasa	Serial drama	A Spanish dramedy streaming television series. Set in Seville, the fiction follows the life of Pedro Marrero, aka "El Grasa", an overweight criminal with an unhealthy lifestyle who suffers from a heart attack and then decides to radically change his life in order to improve on his health condition and stay alive.
Informativos UMATIC	News show	A mix of interviews and cultural magazine broadcast in the 90s in the territorial center of Aragón.
Jara y Sedal	Thematic magazine	Show dedicated to the world of hunting and fishing that takes place in Spain.
Noticias Nacional	News show	News clips broadcast since the 1960 s.
Riders	Serial drama	A fiction series with a mix of thriller, comedy and elements of social drama.
Saber y Ganar	Game show	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.
Toros	Thematic magazine	Broadcast and raw footage of a TVE show related to bulls.
Yreal	Serial drama	Action thriller television series which blends live-action footage with 2D animation.

**Table 3.** RTVE2022DB test dataset distribution among the different challenges. S2T: Speech to Text, SDIA: Speaker Diarization and Identity Assignment, TaSA: Text and Speech Alignment, SoS: Search on Speech.

TV Show Name	RTVE2022DB				
	Show Code	S2T	SDIA	TaSA	SoS
3 × 4	3 × 4	2:58:17	2:58:17		0:59:21
A pedir de boca	APB	3:41:38			
Agrosfera	AG	4:15:20	4:15:20	6:38:04	0:51:27
Aquí la Tierra	AT	2:46:44	2:46:44	2:52:35	0:28:51
Ateneo	ATE	1:40:09			
Cerámica Popular Española	CPE	1:02:35			
Comando Actualidad	CA	3:59:29	3:59:29		1:00:39
Conversatorios en Casa América	CCA	1:58:44			
Corazón	CO	3:00:17	3:00:17	4:33:49	0:30:03
El cazador	EC	3:48:22			
Encuestas con ruido ambiente	ERA	2:08:13			
Entrevistas en estudio	EE	3:54:57			
España Directo	ED	4:05:57	4:05:57		0:59:01
Grasa	GR	1:29:36	1:29:36		
Informativos UMATIC	IU	0:59:49			
Jara y Sedal	JyS	2:29:17			
Noticias Nacional	NN	2:14:32			
Riders	RD	1:15:00	1:15:00		
Saber y Ganar	SyG	4:28:28			
Toros	TO	0:49:57			
Yreal	YR	1:08:46	1:08:46		
Total hours		54:16:07	24:59:26	14:04:28	4:49:22

## 2.2. MAVIR Database

The MAVIR database consists of a set of Spanish talks from the MAVIR workshops (<http://www.mavir.net> (accessed on 24 July 2023)) held in 2006, 2007 and 2008. It contains speech samples from Spanish speakers both from Spain and Latin America.

The MAVIR Spanish dataset consists of 7 h of spontaneous speech files from different speakers. These data were then divided into training, development, and test sets. The speech data were manually annotated in an orthographic form, but timestamps were only set for phrase boundaries (<http://cartago.llf.uam.es/mavir/index.pl?m=videos> (accessed on 24 July 2023)). The training data were made available to the participants including the orthographic transcription and the timestamps for phrase boundaries. For the challenge, the timestamps for the roughly 3000 occurrences of the queries used in the development and test evaluation datasets were also provided.

Initially, the speech data were recorded in several audio formats (pulse code modulation (PCM) mono and stereo, MP3, 22.05 kHz, and 48 kHz, among others). For this evaluation, audio recordings were all converted to PCM, 16 kHz, single channel, 16 bits per sample using the SoX tool (<http://sox.sourceforge.net/> (accessed on 24 July 2023)). All the recordings but one were originally made with a Digital TASCAM DAT model DA-P1 equipment. Different microphones were used, which mainly consisted of tabletop or floor-standing microphones, but one lavalier microphone was also employed. The distance from the microphone to the mouth of the speaker was not specifically controlled, but in most cases was smaller than 50 cm. The speech recordings took place in large conference rooms with a capacity for over a hundred people. This conveys additional challenges

including background noise (particularly babble noise) and reverberation. Therefore, these realistic settings and the variety of phenomena in spontaneous speech make this database appealing and challenging enough for evaluation.

A summary of the main database features such as the training/development/test dataset division, the number of word occurrences, the duration, and the number of speakers is presented in Table 4.

The number of terms and the number of total occurrences both for STD (Spoken Term Detection) and QbE STD (Query-by-Example Spoken Term Detection) tasks for the MAVIR database are presented in Table 5.

**Table 4.** MAVIR database: number of word occurrences (#occ.), duration (dur.) in minutes (min.), and number of speakers (#spk.) for training (train), development (dev) and testing (test) datasets.

File ID	Data Type	#occ.	dur. (min)	#spk.
Mavir-02	train	13,432	74.51	7 (7 males)
Mavir-06	train	4332	29.15	3 (2 males, 1 female)
Mavir-08	train	3356	18.90	1 (1 male)
Mavir-09	train	11,179	70.05	1 (1 male)
Mavir-12	train	11,168	67.66	1 (1 male)
Mavir-03	dev	6681	38.18	2 (1 male, 1 female)
Mavir-07	dev	3831	21.78	2 (2 males)
Mavir-04	test	9310	57.36	4 (3 males, 1 female)
Mavir-11	test	3130	20.33	1 (1 male)
Mavir-13	test	7837	43.61	1 (1 male)
ALL	train	43,467	260.67	13 (12 males, 1 female)
ALL	dev	10,512	59.96	4 (3 males, 1 female)
ALL	test	20,277	121.3	6 (5 males, 1 female)

**Table 5.** MAVIR database query information: number of terms and number of occurrences for STD and QbE STD tasks both for development and test datasets ('dev.' stands for development). The term length of the development query lists varies between 4 and 27 graphemes. The term length of the MAVIR test query lists varies between 4 and 28 graphemes.

Task	#dev. Terms	#dev. Occurrences	#test Terms	#test Occurrences
STD	374	1014	223	2121
QbE STD	102	425	106	1192

### 2.3. RTVE2022-SoS Database

An excerpt of the whole RTVE database presented in Table 1 has been used as development data for the search on speech challenge. Specifically, the search on speech challenge provided two different development datasets. The dev1 dataset consists of about 60 h of speech with human-revised word transcriptions without time alignment. The dev2 dataset, which was the one actually used as development dataset for the search on speech challenge, consists of 15 h of speech data. For the challenge, the timestamps for the roughly 2500 occurrences of the queries used in the development (dev2) and test evaluation datasets were also provided.

A summary of the main database features such as the development (dev2)/test dataset division, the number of word occurrences, the duration, and the number of speakers is presented in Table 6.

The number of terms and the number of total occurrences both for STD and QbE STD tasks for the RTVE2022-SoS database are presented in Table 7.



**Table 6.** RTVE2022-SoS database: number of word occurrences (#occ.), duration (dur.) in minutes (min.), and number of speakers (#spk.) development (dev2) and testing (test) datasets.

File ID	Data Type	#occ.	dur. (min)	#spk.
LN24H-20151125	dev2	21,049	123.50	22
LN24H-20151201	dev2	19,727	112.43	16
LN24H-20160112	dev2	18,617	110.40	19
LN24H-20160121	dev2	18,215	120.33	18
millennium-20170522	dev2	8330	56.50	9
millennium-20170529	dev2	8812	57.95	10
millennium-20170626	dev2	7976	55.68	14
millennium-20171009	dev2	9863	58.78	12
millennium-20171106	dev2	8498	59.57	16
millennium-20171204	dev2	9280	60.25	10
millennium-20171211	dev2	9502	59.70	12
millennium-20171218	dev2	9386	55.55	15
DG00090476	test	9683	60.65	52
DG90266390	test	8127	59.37	21
DG90715506	test	4533	30.06	46
DG90721106	test	4966	28.87	22
DG90734223	test	9518	59.03	55
I00920573	test	8397	51.46	77
ALL	dev2	149,255	930.64	173
ALL	test	45,224	289.44	273

**Table 7.** RTVE2022-SoS database query information: number of terms and number of occurrences for STD and QbE STD tasks both for development and test datasets ('dev.' stands for development, specifically for the dev2 dataset). The term length of the development query lists varies between 4 and 25 graphemes. The term length of the RTVE test query lists varies between 4 and 27 graphemes.

Task	#dev. Terms	#dev. Occurrences	#test Terms	#test Occurrences
STD	398	1502	260	1039
QbE STD	103	574	107	896

#### 2.4. SPARL22 Database

The SPARL22 database consists of spontaneous speech from Spanish parliament sessions held from 2016 up to now and amounts to about 2 h of speech extracted from 14 audio files. The timestamps for the roughly 1600 occurrences of the queries used as test data were also provided.

The original recordings are videos in MPEG format. The evaluation organizers extracted the audio from these videos and converted them to PCM, 16 kHz, single channel and 16 bits per sample using the ffmpeg 4 tool (<https://ffmpeg.org/> (accessed on 24 July 2023)). It is worth mentioning that this database contains several noise types (e.g., laughing, applause, etc.), which makes it quite challenging.

This database only provides test data to measure the generalization capability of the systems in an unseen domain in training and development.

A summary of the main database features such as the number of word occurrences, the duration, and the number of speakers is presented in Table 8. The number of terms and the number of total occurrences both for STD and QbE STD tasks for the SPARL22 database are presented in Table 9.

**Table 8.** SPARL22 database: number of word occurrences (#occ.), duration (dur.) in minutes (min.), and number of speakers (#spk.) for the testing (test) dataset.

File ID	#occ.	dur. (min)	#spk.
13_000500_003_1_19421_642906	875	5.55	2 (1 male, 1 female)
13_000400_007_0_19432_643097	563	3.53	2 (1 male, 1 female)
13_000400_005_0_19422_642932	718	3.57	2 (1 male 1 female)
13_000400_005_0_19422_642923	1898	11.62	1 (1 female)
13_000400_005_0_19422_642922	1733	11.67	1 (1 female)
13_000400_004_0_19388_642448	1107	7.43	1 (1 male)
13_000400_003_0_19381_642399	1403	8.13	3 (2 males, 1 female)
13_000400_003_0_19381_642398	1279	11.45	3 (2 males, 1 female)
13_000400_002_1_19376_642375	2007	13.70	2 (1 male, 1 female)
13_000400_002_1_19376_642366	1720	10.73	1 (1 male)
13_000327_002_0_19437_643241	1405	8.73	2 (2 males)
12_000400_153_0_18748_633006	1331	8.33	2 (2 males)
12_000400_148_0_18727_632388	1012	5.42	2 (1 male, 1 female)
12_000400_003_0_16430_586456	1484	10.33	1 (1 ma.)
ALL	18,535	120.19	25 (16 males, 9 females)

**Table 9.** SPARL22 database query information: number of terms and number of occurrences for STD and QbE STD tasks for the test dataset. The term length of the SPARL22 test query lists varies between 3 and 26 graphemes.

Task	#test Terms	#test Occurrences
STD	282	1603
QbE STD	108	969

### 2.5. The Basque Parliament Dataset

For the second task of the Text-and-Speech Alignment Challenge (TaSAC-BP), the audio data and the paired texts were extracted from a plenary session of the Basque Parliament (BP), including sections in two languages (Spanish and Basque). The paired texts were extracted from session minutes and included only sections in Spanish. The task focused on Spanish because most of the research groups aiming to participate in this evaluation would have ASR technology and resources for Spanish, but few would have them available for Basque.

#### 2.5.1. Audio Data

The Basque Parliament audio data were stored in 16 kHz 16-bit signed single-channel PCM WAV files. Audio recordings were originally made through the BP audio system (desktop microphones), thus they are generally clear with high signal-to-noise ratios. Two different audio files (each approximately one hour long) were provided for development and testing, respectively. Both audios were extracted from the same plenary session, which featured speech samples from several (not many) speakers, who may switch from Spanish to Basque (or vice versa) during their turns. Speaker turn changes and voting events were both managed by the president of the BP and involved a certain amount of silent or slightly noisy regions, but speaker turn overlaps were very uncommon.

#### 2.5.2. The Paired Texts

The text to be aligned with the audio was extracted from the session's minutes. The Basque Parliament session minutes are based on the audio but ignore spontaneous speech events (such as filled pauses, false starts, repeated words, etc.) and include a sizable amount of editions to preserve syntactical correctness. As a consequence of this, the provided

text does not match the audio, featuring word deletions, insertions and replacements. Sometimes, a word said in the audio is replaced in the minutes with a very similar variation of it (with different gender or number) and the most optimistic alignment will inescapably lead to an error, just because acoustics and spelling do not match. Both the paired texts and the ground truth transcriptions have been normalized by removing punctuation marks, replacing accented vowels with non-accented vowels and converting all letters to lowercase. Uppercase letters have been kept only for acronyms (e.g., ADN, EH, UPyD, etc.) which could be either spelled (the most common case) or read as words. This should be taken into account when performing the alignment.

The paired texts do not include the parts spoken in Basque, so there could be remarkable time leaps between one word, at the end of a part spoken in Spanish, and the following one, at the beginning of the next part spoken in Spanish (maybe several minutes ahead in the audio signal). Again, this should be taken into account when performing the alignment.

### 2.5.3. The Ground Truth

The ground truth is based on manually generated rich text transcriptions (including spontaneous speech events). These transcriptions follow the acoustics even though the syntactical correctness is lost. The timestamps of sentences were manually added, so they are fully reliable. Word-level timestamps inside sentences were obtained automatically by forced alignment of each sentence transcription with the corresponding audio. To verify the accuracy of word-level timestamps, an informal test was carried out using several randomly chosen sentences, by manually adding word-level timestamps and comparing them with the automatically generated timestamps. It was observed that differences between manual and automatic timestamps spanned from 0 to 20 ms. Thus, the automatic segmentation was considered good enough for the purposes of this evaluation, provided that a reasonable collar time was applied.

For this evaluation, only the words appearing in the paired text are kept in the ground truth, the remaining elements of the rich text transcription being hidden. Note that the evaluation focuses on how well the paired text is aligned with the audio. Taking this into account, if a word *w* in the paired text is aligned with an audio segment that is not included in the ground truth, it is guaranteed that neither the word *w* nor any other word in the paired text appears in that segment, so the time span of such segment is counted as error no matter the exact transcription of it. Also, to be fair with the participants, if a word *w* of the paired text (e.g., a proper name) appears in a part of the audio spoken in Basque, the corresponding segment is included in the ground truth, just to cover the case that the word *w* was aligned with that segment. Finally, to account for the uncertainty when defining the borders between words, a collar time can be established so that a certain amount of time around the borders is not evaluated at all.

## 3. IberSpeech-RTVE 2022 Evaluation Tasks

This section presents a brief summary of the four evaluation tasks. A more detailed description of the evaluation plans can be found on the IberSpeech2022 web page ([https://iberspeech2022.ugr.es/?page\\_id=67](https://iberspeech2022.ugr.es/?page_id=67) (accessed on 24 July 2023)) and the Cátedra RTVE-UZ web page (<http://catedrartve.unizar.es/albayzin2022.html> (accessed on 24 July 2023)).

### 3.1. Speech to Text Challenge

The speech-to-text transcription evaluation consists of automatically transcribing different types of TV shows. The main objective is to evaluate the state-of-the-art in automatic speech recognition (ASR) for the Spanish language in the broadcast sector. There is no specific training partition, thus participants are free to use the previous RTVE datasets (2018 and 2020) or any other data to train their systems provided that these data are fully documented in the system's description paper. For public databases, the name of the database must be provided. For private databases, a brief description of the origin of the

data must be provided. Each participant team should submit at least a primary system, but they could also submit up to three contrastive systems.

### Performance Metric

The ASR system outputs are ranked by the word error rate (WER). All the participants have to provide as output for evaluation a free-form text file per test file, encoded using the UTF-8 charset (<http://www.utf-8.com/> (accessed on 24 July 2023)), with no page, paragraphs, sentence, or speaker breaks. The text is normalized as follows: all the punctuation marks are removed, numbers are written with letters and all characters are lower-cased. The WER is defined as:

$$WER = \frac{S + D + I}{N_r}, \quad (1)$$

where  $N_r$  is the total number of words in the reference transcription,  $S$  is the number of substituted words in the automatic transcription,  $D$  is the number of words from the reference deleted in the automatic transcription, and  $I$  is the number of words inserted in the automatic transcription not appearing in the reference.

### 3.2. Speaker Diarization and Identity Assignment Challenge

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, the evaluation asked for assigning the name of these people to the correct diarization labels. No prior knowledge is provided about the number of speakers participating in the audio to be analyzed. Participants are free to use any dataset for training their diarization systems provided that these data were fully documented in the system's description paper. The organization provides all the RTVE datasets (<http://catedrartve.unizar.es/rtvedatabase.html> (accessed on 24 July 2023)), the Catalan broadcast news database from the 3/24 TV channel proposed for the 2010 Albayzin Audio Segmentation Evaluation [15,16] and the Corporación Aragonesa de Radio y Televisión (CARTV) database proposed for the 2016 Albayzin Speaker Diarization evaluation. For the identity assignment task, the RTVE2022 dataset provides enrolment audio files for 74 speakers (38 male, 36 female). At least 30 s of speech from each speaker to identify are included in the dataset.

#### 3.2.1. Diarization Scoring

As in the NIST RT Diarization evaluations (<https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation> (accessed on 24 July 2023)), to measure the performance of the proposed systems, the diarization error rate (DER) is computed as the fraction of speaker time that is not correctly attributed to that specific speaker. This score is computed over the entire file to be processed, including regions where more than one speaker is present (overlap regions).

Given the dataset to evaluate  $\Omega$ , 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:

$$E(n) = T(n) \left[ \max(N_{ref}(n), N_{sys}(n)) - N_{Correct}(n) \right] \quad (2)$$

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

$$DER = \frac{\sum_{n \in \Omega} E(n)}{\sum_{n \in \Omega} (T(n) N_{ref}(n))}. \quad (3)$$

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.
- Missed speech time: The missed speech time refers to the amount of time that speech is present but not labeled by the diarization system.
- 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.

Consecutive speech segments of audio labelled with the same speaker identification tag and separated by a non-speech segment less than 2 s long are merged and considered a single segment. A region of 0.25 s around each segment boundary, usually known as the forgiveness collar, is considered. These regions are excluded from the computation of the diarization error in order to take into account both inconsistent human annotations and the uncertainty about when a speaker turn begins or ends.

### 3.2.2. Identity Assignment Scoring

For the Identity Assignment Task, the assignment error rate (AER) is used, which is a slightly modified version of the previously described DER. 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:

$$AER = \frac{FA + MISS + SPEAKER ERROR}{REFERENCE LENGTH} \quad (4)$$

where:

- *FA* represents the False Alarm Time, which contains the length of the silence segments or 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* (Speaker Error Time) considers the length of the speech segments that belong to speakers of interest attributed to an incorrect speaker.
- *REFERENCE LENGTH* is the sum of the lengths of all the speech segments uttered by the people of interest (i.e., those identities for which the participants will have audio to train their models).

### 3.3. Text and Speech Alignment Challenge

The Text and Speech Alignment Challenge (TaSAC) consists of two subchallenges:

- Alignment of re-speaking generated subtitles (TaSAC-ST). This challenge consists of synchronizing the broadcast subtitles created by re-speaking different TV shows.
- Alignment and validation of speech signals with partial and inaccurate text transcriptions (TaSAC-BP). This challenge consists of aligning text and audio extracted from a plenary session of the Basque Parliament.

#### 3.3.1. Alignment of Re-Speaking Generated Subtitles (TaSAC-ST)

The IberSPEECH-RTVE 2022 Text and Speech Alignment Challenge aims to evaluate the text-to-speech alignment systems on the actual problem of synchronizing re-speaking subtitles with the corresponding audio. The task assesses the state of the art of offline alignment technology. The purpose is to provide subtitles without delay for a new broadcast. In this task, participants are supplied with the subtitles as they originally appeared on TV, including the start and end timestamps of each subtitle. Participants must provide an output with the exact same sequence of subtitles but with new start and end timestamps for each subtitle. It should be noted that re-speaking subtitles often differ from the actual

spoken words. If the speech is too fast the re-speaker tends to suppress words (deletions) or even to paraphrase, which introduces a new level of difficulty in the alignment process. The performance is measured by computing the time differences between the aligned start and end timestamps given by the alignment systems and the reference timestamps derived from a careful manual alignment.

#### Performance Metric

The average program time-error metric (APTEM) is the primary metric for the Text and Speech alignment task. For each program in the test, a program time-error metric (PTEM) will be calculated and the final score will be computed by averaging the PTEM of each program in the test.

$$APTEM = \frac{1}{M} \sum_{m=1}^M PTEM(m), \quad (5)$$

where  $M$  is the number of programs in the test dataset, and the PTEM is computed as follows: Given the probability distribution of the time difference between the reference and aligned start and end timestamps of each subtitle in a program, the PTEM is computed as the median value of the distribution.

Let's define the start-time error for the  $n$ -th subtitle,  $TE_s(n)$ , as:

$$TE_s(n) = |Ts_{ref}(n) - Ts_{align}(n)|, \quad (6)$$

where  $Ts_{ref}(n)$  and  $Ts_{align}(n)$  are the start timestamps of the reference and the automatic alignment for the  $n$ -th subtitle. Similarly, the end-time error,  $TE_e(n)$ , is defined as:

$$TE_e(n) = |Te_{ref}(n) - Te_{align}(n)|, \quad (7)$$

where  $Te_{ref}(n)$  and  $Te_{align}(n)$  are the end timestamps of the reference and the automatic alignment for the  $n$ -th subtitle. Then, the time error of the  $n$ -th subtitle,  $TE(n)$  is computed as:

$$TE(n) = TE_s(n) + TE_e(n). \quad (8)$$

The PTEM is defined as

$$PTEM = med([TE(1), TE(2), \dots, TE(N)]), \quad (9)$$

where  $med()$  is the median operator and  $N$  is the number of subtitles broadcast in the TV program.

#### 3.3.2. Alignment and Validation of Speech Signals with Partial and Inaccurate Text Transcriptions (TaSAC-BP)

Over the last years, with the widespread adoption of data-intensive deep learning approaches to ASR, the semi-supervised collection of training data for ASR has gained renewed interest. The Internet is plenty of resources pairing speech and text. Sometimes the paired text is an accurate transcription of the spoken content, but frequently it is only a loose or partial transcription or even a translation to some other language. Therefore, a text-to-speech alignment system able to detect and extract accurately paired speech and text segments becomes a very valuable tool. The second task of the Text-and-Speech Alignment Challenge (TaSAC-BP) was designed with that goal in mind. The alignment systems would deal with a long audio file, including sections in Spanish and Basque, but the paired texts (which could be partial or approximate transcripts of the audio) would cover only the Spanish sections. The audio parts in Basque were not expected to be paired with any text, though some words or word fragments (proper names, technical terms, etc.) may actually match (and be wrongly paired with) text in Spanish.

### Task Description

The task consisted of aligning each word of the text with a segment of the audio file so that the audio content corresponds to the pronunciation of the given word. Alignments were required to be monotonous, that is, the sequence of timestamps had to be non-decreasing. Obviously, it was guaranteed that there was an optimal monotonous alignment between the audio signal  $X$  and the paired text  $W$ . Let  $W = \{w_1, w_2, \dots, w_N\}$  be the sequence of  $N$  words to be aligned with an audio signal  $X$ , and let  $S = \{s_1, s_2, \dots, s_N\}$  be the corresponding sequence of aligned segments in  $X$ . Then, if a word  $w_i$  is aligned to a segment  $s_i = (t_{beg}^{(i)}, t_{end}^{(i)})$  and another word  $w_j$  is aligned to a segment  $s_j = (t_{beg}^{(j)}, t_{end}^{(j)})$ , with  $i < j$ , then the timestamps defining those segments must be  $t_{beg}^{(i)} < t_{end}^{(i)} \leq t_{beg}^{(j)} < t_{end}^{(j)}$ . Non-monotonic alignments were not allowed and non-monotonic submissions were not accepted.

The output of an alignment system was required to be a text file containing a line for each word in the paired text, each line including 5 columns (separated by any amount of spaces or tabs) with the following information:

- $t_{beg}$ : A real number with the time when the segment starts.
- $t_{end}$ : A real number with the time when the segment ends.
- **word**: The word paired with the audio segment.
- **score**: A real number reflecting the confidence on the alignment, the more positive the score, the higher the confidence; the more negative the score, the lower the confidence.
- **decision**: A binary value (0/1), 0 meaning Reject and 1 meaning Accept. Since only the accepted words would be evaluated, this decision should be made by applying a confidence score threshold.

The participants could submit results for at most five (one primary + four contrastive) systems. Each system should automatically align the paired text with the audio, taking into account that some parts of the audio should not be aligned with any text and that the paired text did not reflect exactly the audio contents. It was not allowed to listen to the audio or use any kind of human intervention (e.g., crowdsourcing). Otherwise, any approach could be applied with no limit to the type or amount of resources that the participants could use to perform the task, as long as the employed methods and resources were described with enough detail and, if possible, links to papers, data and/or software repositories were provided to make it easier to reproduce their approach. For each system, two separate result files were required, for the development and test sets, respectively. Finally, participants would be ranked according to the performance obtained by their primary systems on the test set.

### Performance Metric

The ground truth is preprocessed before using it to compute the performance metric. First, the missing segments are added to the ground truth and assigned an out-of-vocabulary label (#). Then, the borders between segments are redefined by excluding from evaluation a collar time  $t_{collar}$  around them (in this evaluation,  $t_{collar} = 20$  ms): the starting time  $t_{beg}$  of each segment is added  $t_{collar}/2$  while the ending time  $t_{end}$  of each segment is subtracted  $t_{collar}/2$ .

Since the objective of the alignment is to recover as much correctly transcribed speech as possible to train acoustic models for the development of ASR systems, the performance metric should reflect this objective, but also the negative impact of wrongly aligned segments that could seriously compromise this semi-supervised training strategy. Thus, the performance metric will be just the difference between the correct and the wrongly aligned times.

Let  $S = \{s_1, s_2, \dots, s_N\}$  be the output of the alignment system for the paired text  $W = \{w_1, w_2, \dots, w_N\}$ . Only those segments accepted by the system will be evaluated, so let  $S' = \{s_1, s_2, \dots, s_{N'}\}$  (with  $N' \leq N$ ) be the sequence of accepted segments. Each accepted segment is then aligned with the ground truth, which produces a sequence of

sub-segments, each of them aligned either with a ground truth segment or with a collar time segment.

Sub-segments aligned with collar time are not evaluated and will not be considered hereafter. Let  $C = \{c^{(1)}, c^{(2)}, \dots, c^{(M)}\}$  be the sequence of sub-segments obtained after aligning the accepted segments with the ground truth, excluding collar time. Each sub-segment is a 4-tuple:

$$c^{(i)} = (t_{beg}^{(i)}, t_{end}^{(i)}, w_a^{(i)}, w_g^{(i)}), \quad (10)$$

where  $t_{beg}^{(i)}$  is the starting time,  $t_{end}^{(i)}$  is the ending time,  $w_a^{(i)}$  is a word in the paired text and  $w_g^{(i)}$  is a word in the ground truth. The *performance metric* is defined as follows:

$$score(C) = \sum_{i=1}^M \delta(w_a^{(i)}, w_g^{(i)}) \cdot (t_{end}^{(i)} - t_{beg}^{(i)}), \quad (11)$$

where:

$$\delta(w_a^{(i)}, w_g^{(i)}) = \begin{cases} 1 & \text{if } w_a^{(i)} = w_g^{(i)} \\ -1 & \text{if } w_a^{(i)} \neq w_g^{(i)}. \end{cases} \quad (12)$$

The participants were encouraged to take into account that the higher the number of accepted segments, the higher the potential amount of correctly aligned speech, but also the higher the risk of having a large amount of wrongly transcribed speech. To find the optimal balance between both events, a suitable confidence score threshold should be applied to make decisions. The scoring script provided by the organization explores all the possible thresholds that can be applied to make decisions, and outputs the optimal score and threshold.

### Scoring Script

The scoring script provided to participants is a command-line application that requires a basic installation of Python 3 including the matplotlib module (used to produce a graphical analysis of system scores). The script takes the system alignment and the ground truth files as input and allows to specify collar time (in this case, 0.02 s) as well as other optional arguments, such as the text and graphical output file names.

The output text includes two lines, the first one showing the performance obtained using system decisions, and the second one showing the best performance that can be obtained by applying a threshold on the provided scores to make decisions. By default, the text output is written on the console. The optional graphical output (a PNG file) presents the performance obtained by applying system decisions and the evolution of the correctly aligned time, the wrongly aligned time and the difference between them (that is, the performance metric) by using all the possible thresholds to make decisions. The optimal performance and the corresponding threshold are marked on the performance curve. The figure also includes the total time accepted and rejected by applying different thresholds. Obviously, applying the minimum threshold implies accepting all the words of the paired text, which does not usually yield the best performance, while applying the maximum threshold implies rejecting all the words, meaning a performance of 0. A reasonable criterion to make decisions on the test set would be to apply the optimal threshold found on the development set.

### 3.4. Search on Speech Challenge

The Search on Speech challenge involves searching in audio content a list of terms or queries and it is suitable for groups working on speech indexing/retrieval and speech recognition. In other words, this challenge focuses on retrieving the audio files that contain any of those terms/queries along with the corresponding timestamps.

This challenge consists of two different tasks:



- Spoken Term Detection (STD), where the input to the system is a list of terms, but these terms are unknown when processing the audio. This task must generate a set of occurrences for each term detected in the audio files, along with their timestamps and score as output. This is the same task as in NIST STD 2006 evaluation [17] and Open Keyword Search in 2013 [4], 2014 [5], 2015 [6], and 2016 [7].
- Query-by-Example Spoken Term Detection (QbE STD), where the input to the system is an acoustic query and hence a prior knowledge of the correct word/phone transcription corresponding to each query cannot be used. This task must generate a set of occurrences for each query detected in the audio files, along with their timestamps and score as output, as in the STD task. This QbE STD is the same task as those proposed in MediaEval 2011, 2012, and 2013 [18].

For the QbE STD task, participants are allowed to make use of the target language information (Spanish) when building their system/s (i.e., system/s can be language-dependent). Nevertheless, participants are strongly encouraged to build language-independent QbE STD systems, as in past MediaEval Search on Speech evaluations, where no information about the target language was given to participants.

This evaluation defined two different sets of terms/queries for STD and QbE STD tasks: an in-vocabulary (INV) set of terms/queries and an out-of-vocabulary (OOV) set of terms/queries from lexicon and language model perspectives. The OOV set of terms/queries will be composed of out-of-vocabulary words for the LVCSR system. This means that, in case participants employ an LVCSR system for processing the audio for any task (STD, QbE STD), these OOV terms (i.e., all the words that compose the term) must be previously removed from the system dictionary/language model and hence, other methods (e.g., phone-based systems) have to be used for searching OOV terms/queries. Participants can consider OOV words for acoustic model training if they find it suitable.

Regarding the QbE STD task, three different acoustic examples per query were provided for both development and test datasets. One example was extracted from the same dataset as the one to be searched (hence in-domain acoustic examples). This scenario considered the case in which the user finds a term of interest within a certain speech dataset and he/she wants to search for new occurrences of the same query. The two other examples were recorded by the evaluation organizers and comprised a scenario where the user pronounces the query to be searched (hence out-of-domain acoustic examples). These two out-of-domain acoustic examples amount to 3 s of speech with PCM, 16 kHz, single channel and 16 bits per sample with the microphone of an HP ProBook Core i5, 7th Gen and with a Sennheiser SC630 USR CTRL microphone with noise cancellation, respectively.

The queries employed for the QbE STD task were chosen from the STD queries. It should be noted that for both the STD and QbE STD tasks, a multi-word query was considered OOV in case any of the words that form the query was OOV.

### Evaluation Metric

In search on speech systems (both for STD and QbE STD tasks), a hypothesized occurrence is called a detection; if the detection corresponds to an actual occurrence, it is called a hit; otherwise, it is called a false alarm. If an actual occurrence is not detected, this is called a miss. The main metric for the evaluation is the actual term weighted value (ATWV) metric proposed by NIST [17]. This metric combines the hit rate and false alarm rate of each query and averages over all the queries, as shown in Equation (13):

$$ATWV = \frac{1}{|\Delta|} \sum_{Q \in \Delta} \left( \frac{N_{hit}^Q}{N_{true}^Q} - \beta \frac{N_{FA}^Q}{T - N_{true}^Q} \right), \quad (13)$$

where  $\Delta$  denotes the set of queries and  $|\Delta|$  is the number of queries in this set.  $N_{hit}^Q$  and  $N_{FA}^Q$  represent the numbers of hits and false alarms of query  $Q$ , respectively and  $N_{true}^Q$  is the number of actual occurrences of query  $Q$  in the audio.  $T$  denotes the audio length in

seconds and  $\beta$  is a weight factor set to 999.9, as in the ATWV proposed by NIST [17]. This weight factor causes an emphasis placed on recall compared to precision with a 10:1 ratio.

ATWV represents the term weighted value (TWV) for an optimal threshold given by the system (usually tuned on the development data). An additional metric, called maximum term weighted value (MTWV) [17] is also used to evaluate the upper-bound system performance regardless of the decision threshold.

Additionally,  $p(\text{Miss})$  and  $p(\text{FA})$ , which represent the probability of miss and false alarm of the system as defined in Equations (14) and (15), respectively, are also reported:

$$p(\text{Miss}) = 1 - \frac{N_{hit}}{N_{true}} \quad (14)$$

$$p(\text{FA}) = \frac{N_{FA}}{T - N_{true}}, \quad (15)$$

where  $N_{hit}$  is the number of hits obtained by the system,  $N_{true}$  is the actual number of occurrences of the queries in the audio,  $N_{FA}$  is the number of false alarms produced by the system and  $T$  denotes the audio length (in seconds). These values, therefore, provide a quantitative way to measure system performance in terms of misses (or equivalently, hits) and false alarms.

In addition to ATWV, MTWV,  $p(\text{Miss})$  and  $p(\text{FA})$  figures, NIST also proposed a detection error tradeoff (DET) curve [19] that evaluates the performance of a system at various miss/FA ratios. Although DET curves were not used for the evaluation itself, they are also presented in this paper for system comparison.

The NIST STD evaluation tool [20] was employed to compute the MTWV, ATWV,  $p(\text{Miss})$  and  $p(\text{FA})$  figures, along with the DET curves.

## 4. IberSpeech-RTVE 2022 Submitted Systems

### 4.1. Speech to Text Challenge

A total of 13 different systems from four participating teams were submitted. The most relevant characteristics of each system are presented in terms of the recognition engine, and audio and text data used for training acoustic and language models.

- BCN2BRNO** [21]. BUT Speech@FIT research group (Brno University of Technology, Czech Republic) and Telefónica Research (Spain)
 

BCN2BRNO submitted a primary system based on a word-level ROVER fusion of five individual models. Three models used an Encoder-Decoder Transformer architecture (XLS-R Conformer, XLSR-53-CTC and Whisper large model), the fourth was an RNN Transducer architecture and the fifth was a hybrid DNN-HMM model. The ASR pipeline consisted of 4 conceptual blocks: (a) voice activity detection to split audio recordings into smaller segments, (b) ASR models to produce N-best lists of hypotheses and scores, (c) RNN-LM rescoring to produce 1-best hypothesis for each ASR system, and (d) shallow fusion to give a single output. The system achieved 13.7% WER on the official evaluation dataset in the previous Albayzin 2020 Speech to Text Transcription challenge where the best result was 16.04% WER. The data used for training the ASR models contained the RTVE2018, 2020 and 2022 databases and the Spanish Common Voice dataset (400 h of read speech). Noise data augmentation was used with the Spanish Common Voice dataset to produce an augmented data partition with an SNR range from 6 dB to 20 dB and different noises (restaurant, street, home, workshop, fan, and non-speech segments longer than 2 s from the RTVE dataset). The language model used for rescoring was trained on a collection of Spanish newspapers' texts and fine-tuned to the transcripts of the training data. Three contrastive systems were submitted: (c1) the fusion of an XLSR-128-Conformer with data augmentation and a CNN-TDNN-HMM with x-vectors; (c2) a single system based on XLSR-128-Conformer, and (c3) the Whisper large model.
- UZ** [22]. ViVoLab research group (University of Zaragoza, Spain)

The ViVoLab system was a combination of several subsystems designed to perform a full subtitle edition process from the raw audio to the creation of aligned subtitle transcribed partitions. The subsystems included a phonetic recognizer, a phonetic subword recognizer, a speaker-aware subtitle partitioner, a sequence-to-sequence translation model working with orthographic tokens to produce the desired transcription, and an optional diarization step with the previously estimated segments. Additionally, the system used recurrent network-based language models to improve results for steps that involve search algorithms like the subword decoder and the sequence-to-sequence model. The technologies involved include unsupervised models like Wavlm to deal with the raw waveform, convolutional, recurrent, and transformer layers. The acoustic models were trained using the following datasets: (a) Albayzin, a phonetically balanced corpus in Spanish with 12 h, (b) Speech-Dat-Car, a corpus recorded in a car in different driving conditions with 18 h, (c) Domolab, a corpus recorded in a domestic environment with 9 h, (d) TCSTAR transcriptions of Spanish parliament sessions with 111 h, (e) Common Voice in Spanish with 400 h and (f) RTVE 2018 train, dev and RTVE 2022 train with more than 800 h. In addition, more training material was added from a diverse scarp of online videos and social networks with a total of 10,000 h. These data were filtered after performing forced alignment and selecting transcribed segments with sufficient quality. Data augmentation was used including noises, music and artificially generated impulse responses to simulate different acoustic environments. The language model was trained using Spanish Wikipedia, RTVE challenge subtitles, and text obtained from different Spanish newspapers. The system achieves 21.86% WER on the official evaluation dataset in the previous Albayzin 2020 Speech to Text Transcription challenge, where the best result was 16.04% WER.

- **TID [23].** Telefónica I + D (Spain)

The TID system consisted of an acoustic end-to-end ASR based on the XLSR-53 model pre-trained with 56k hours of audio in 53 languages and fine-tuned in the Spanish Common Voice dataset. A voice activity detection (VAD) model was used to filter non-speech segments. The transcription was obtained using simple greedy decoding from the frame-wise character posteriors given by the model. A self-supervised method was used to adapt the ASR to the TV broadcast domain by retrieving data from the RTVE2018 and 2022 datasets. A neural machine translation model was used as an external language model to correct the ASR hypothesis. The primary system used VAD segments and corrected the output with the language model. Three additional contrastive systems were submitted: (c1) just used the VAD segments, (c2) used 10-second-length window segmentation and corrected the output with the language model, and (c3) generated the output using a 10-s-length segmentation window size.

- **VICOM-UPM [24].** Fundación Vicomtech (Basque Research and Technology Alliance) and Polytechnic University of Madrid (Spain)

VICOM-UPM submission consisted of a total of 4 different systems which allowed testing state-of-the-art modeling approaches focused on different learning techniques and typologies of neural networks. As primary systems, VICOM-UPM presented the pre-trained Wav2Vec2-XLS-R-1B model adapted with 300 h of in-domain data from the RTVE2018 and RTVE2020 datasets. The first contrastive system corresponded to a pruned RNN-Transducer model, composed of a Conformer encoder and a stateless prediction network using BPE word-pieces as output symbols. The second contrastive system was a Multistream-CNN acoustic model-based system with a non-pruned 3-gram model for decoding and an RNN-based language model for rescoring the initial lattices. Finally, the third contrastive system was the publicly available Large model of the Whisper engine. The acoustic corpus was composed of annotated audio contents from 9 different datasets, summing up a total of 1927 h. The datasets were: (a) RTVE2018, (b) SAVAS corpus (broadcast news contents in Spanish of the Basque Country's public broadcast corporation), (c) IDAZLE corpus (TV shows from the Basque Country's public broadcast corporation), (d) RTVE Play 2020 and 2022

including programs broadcast between 2018 and 2022 by RTVE, (e) RTVE youtube including contents of RTVE from the YouTube platform, (f) Spanish Common Voice, (g) Albayzin, and (h) Multext (multilingual prosodic database). The text corpus consisted of transcriptions of the training datasets, RTVE2018 subtitles, RTVE play and news subtitles and generic news gathered from digital newspapers on the Internet. With regard to the RTVE 2020 test dataset, the best results were obtained by the Whisper model (12.15% WER) followed by the primary system (13.77%).

#### 4.2. Speaker Diarization and Identity Assignment Challenge

Only 2 teams have participated in the Speaker Diarization challenge and only one of them submitted results with Identity Assignment.

- **IRIT [25].** Université de Toulouse (France)  
The IRIT submission was based on the pyannote.audio diarization system. pyannote.audio (<https://github.com/pyannote/pyannote-audio>) (accessed on 24 July 2023) is an open-source toolkit written in Python for speaker diarization. Version 2.1 introduces a major overhaul of pyannote.audio default speaker diarization pipeline, made of three main stages: speaker segmentation applied to a short sliding window, neural speaker embedding of each (local) speaker, and (global) agglomerative clustering. The IRIT submission focused on Speaker Diarization.
- **IV [26].** Intelligent Voice (UK)  
The Intelligent Voice submission was based on a Variational Bayes x-vector Voice Print Extraction system. The proposed system is capable of capturing vocal variations using multiple x-vector representations with two-stage clustering and outlier detection refinement. It implements the Deep-Encoder Convolution Autoencoder Denoiser network for denoising segments with noise or music on files identified by a signal-to-noise ratio classifier. Intelligent Voice submitted systems for both Speaker Diarization and Identity Assignment.

#### 4.3. Text and Speech Alignment Challenge

Only two participants submitted results to this evaluation, maybe due to a lack of interest, because text and speech alignment is just a preprocessing step in the development of ASR systems and off-the-shelf tools are good enough in most cases.

##### 4.3.1. Alignment of Re-Speaking Generated Subtitles (TaSAC-ST)

Only one team submitted results to this evaluation:

- **GTTS [27].** University of the Basque Country UPV/EHU (Spain)  
This system was originally developed as a baseline for the TaSAC-BP subtask and was adjusted to meet the conditions (output file format, etc.) of TaSAC-ST. Acoustic models were trained on cross-domain (non-RTVE) materials, with different channels, background/environment conditions, speakers, etc. The original GTTS submission consisted of two systems (primary and contrastive-1), which applied the same approach but different acoustic models (trained on two independent sets of non-RTVE data). The primary system used a 332-h training set consisting of clean read speech in Basque and Spanish extracted from generic acoustic databases (including Mozilla Common Voice), while the contrastive-1 system used a 1000-h training set consisting of clean spontaneous speech in Basque and Spanish extracted from Basque Parliament plenary sessions. Two late (post-key) systems (contrastive-2 and contrastive-3, trained on the same datasets as the primary and contrastive-1 systems, respectively) were also submitted to the evaluation, yielding improved performance thanks to a kernel modification in the dynamic programming algorithm which provided more compact alignment hypotheses.  
The four GTTS systems relied on a set of acoustic models used to perform an unrestricted phone decoding of the audio signal, without any language or phonological

models to help the decoding. Given an audio file  $X$  and the corresponding STM file with the subtitles  $S$ , an off-the-shelf end-to-end neural-network-based phone decoder (built with wav2letter++, see [28]) was applied to  $X$ , which produced a recognized sequence of phone-like units  $p_X$  (with timing information attached). On the other hand, an in-house rule- and dictionary-based grapheme-to-phoneme (G2P) converter was applied to  $S$ , which produced a reference sequence of phone-like units  $p_S$  (with word and subtitle information attached). The two sequences of phone-like units were aligned under the criterion of maximizing the number of matches in the alignment path. Finally, the timing information was transferred from  $p_X$  to  $p_S$  and a new STM file was created, identical to the source STM except for the timestamps, which were obtained from the alignment.

#### 4.3.2. Alignment and Validation of Speech Signals with Partial and Inaccurate Text Transcriptions (TaSAC-BP)

For this evaluation, the organizing team developed a baseline system that aimed to establish a reference mark for participants. The baseline system was based on an off-the-shelf phone decoder (built with wav2letter++, see [28]), an in-house Grapheme-to-Phoneme (G2P) converter, and a dynamic programming alignment algorithm maximizing the number of matches in the alignment path. Given an audio  $X$  and the paired text  $S$ , the phone decoder was applied to  $X$  to get a phonetic sequence  $p_X$  (with timestamps attached), the G2P converter was applied to  $S$  to get a second phonetic sequence  $p_S$  (with word info attached), then the two sequences  $p_X$  and  $p_S$  were aligned and the timing information was passed from  $p_X$  to  $p_S$ , and from  $p_S$  to words. Word timestamps were post-processed to fill small gaps between words and the word score was computed as the phone recognition rate in the alignment. Finally, the acceptance threshold was optimized on the development set and applied to the development and test sets.

Only one team submitted results to this evaluation:

- **BUT.** BUT Speech@FIT research group (Martin Kocour and Martin Karafiát, from Brno University of Technology, Brno, Czech Republic)

This submission leveraged the output of some of the ASR systems developed by the BCN2BRNO team for the Speech to Text Challenge [21]. It consisted of a primary system based on a fusion of three ASR systems (two of them based on an encoder-decoder transformer architecture: XLS-R Conformer and Whisper large model, and the third one based on an RNN transducer architecture) and a contrastive system based on the best single ASR system (XLS-R Conformer). The audio processing pipeline started with Voice Activity Detection (VAD), based on a simple feed-forward DNN with 2 outputs (speech/non-speech) trained on the whole RTVE 2018 dataset, which produced segmented audio. Then, ASR was performed on this segmented audio and a 1-best hypothesis was obtained. This 1-best hypothesis was aligned with the original transcript, which produced segmented text, that is, the reference text was sequentially assigned to different audio segments. The alignment process involved filtering out words not recognized by ASR; still, if a word was not found in ASR but its surrounding words (two on each side) were found, the word was kept. Finally, forced alignment was performed between the segmented text and the segmented audio, which produced a sequence of aligned words (words with timestamps). Forced alignment was run on each audio segment separately, with 10-ms steps, using a GMM-HMM-based model (Kaldi) trained on RTVE datasets.

#### 4.4. Search on Speech Challenge

In this challenge, only two systems have been evaluated:

- **Multistream CNN system** [24]. Fundación Vicomtech (Basque Research and Technology Alliance)

This is the same system as the second contrastive system submitted by Vicomtech to

the Speech to Text Challenge, without the RNN language model rescoring approach. For detecting terms, a simple search of the terms within the ASR output produces the detected term list.

- **Multistream CNN+rescoring system** [24]. Fundación Vicomtech (Basque Research and Technology Alliance)

This is the same system as the second contrastive system submitted by Vicomtech to the Speech to Text Challenge. As in the previous system, a simple search of the terms within the ASR output produces the detected term list.

## 5. IberSpeech-RTVE 2022 Results and Discussion

### 5.1. Speech to Text Challenge

Table 10 presents the overall results for the RTVE2022DB test dataset. The more competitive systems have been the ones submitted by BCN2BRNO and VICOM-UPM teams. The best results have been obtained by the BCN2BRNO team, where their primary system (a fusion of 5 individual models) achieves a word error rate of 14.35%. It is worth noting the good results of the VICOM-UPM team, as their first contrastive system based on a pruned RNN-Transducer model obtains a remarkable 14.78% WER. Two contrastive systems are based on the large model of the Whisper engine. Although the data and models are the same, there is a significant difference between the final WER obtained by both submissions. The post-processing strategy of the BCN2BRNO team was based on using *zlib* to compress the segment transcripts and filter out segments with a compression factor higher than 2. The VICOM-UPM team used a more sophisticated post-processing strategy: for transcriptions where more than 20% of segments were text repetitions, the audio was segmented through a VAD module, non-speech segments longer than 2 s were discarded and the decoding was remade. The decoding process of the rest of the speech segments was performed by resetting the pre-condition of the text decoder. In addition, text phrases repeated three times or more were filtered out and only a single appearance was kept. The best WER obtained by a Whisper-based system was 14.87% for the VICOM-UPM approach.

**Table 10.** Speech to Text challenge results for the RTVE2022DB test dataset.

System ID	Primary	C1	C2	C3
BCN2BRNO	14.35%	15.24%	17.22%	18.65%
TID	23.50%	23.45%	24.87%	25.25%
VICOM-UPM	15.30%	14.78%	17.29%	14.87%
UZ	20.32%	26.49%	-	-

It is interesting to study the distribution of WER in terms of the different shows that make up the test set. Table 11 presents the best results by show, that is, for each show the system with the lowest WER has been taken. It is remarkable that 12 out of 21 shows have a WER below 15%, the lowest WER of 5.89% being found for the Agrosfera (AG) show. The three serial dramas, Grasa (GR), Riders (RD) and Yreal (YR) are in the range of 18% to 25% WER. In the Grasa serial drama, most of the players are using Spanish Slang typical of the drug world which makes noteworthy the 24.79% WER. The show EE (indoor interviews) is composed of raw footage of interviews where the cameraman, producer and other people besides the interviewee are talking with a low signal-to-noise ratio, which makes WER rise to 22.2% from the expected 10%. Finally, the worst WER, 47.79%, is obtained for the APB (A Pedir de Boca) show. The audio of this show comes from raw footage where some interviewees are using Spanish Slang, and most of the time people behind the cameras are asking questions or talking, which has made also hard for human annotators to transcribe these materials.

**Table 11.** Speech to Text challenge: Best WER (%) by show. The WER for the 3 best systems (TOTAL WER < 15%) is shown jointly with the best WER among all systems.

Show Code	BCN2BRNO(P)	VICOM-UPM(C1)	WHISPER	BEST WER	Best System
3 × 4	13.11	13.37	14.78	12.60	VICOM-UPM(P)
AG	<b>5.89</b>	6.72	6.16	5.89	BCN2BRNO(P)
APB	49.85	60.24	67.05	47.79	UZ(P)
AT	11.16	11.65	<b>9.60</b>	9.60	Whisper
ATE	9.47	9.07	<b>7.92</b>	7.92	Whisper
CA	<b>17.71</b>	19.09	21.30	17.71	BCN2BRNO(P)
CCA	10.31	<b>9.51</b>	11.93	9.51	VICOM-UPM(C1)
CO	8.55	<b>7.89</b>	9.88	7.89	VICOM-UPM(C1)
CPE	13.87	<b>13.46</b>	14.57	13.46	VICOM-UPM(C1)
EC	13.61	14.32	<b>13.45</b>	13.45	Whisper
ED	14.38	14.21	<b>13.36</b>	13.36	Whisper
EE	23.55	25.16	<b>22.20</b>	22.20	Whisper
ERA	22.08	19.88	<b>18.80</b>	18.80	Whisper
GR	26.63	29.02	31.08	24.79	VICOM-UPM(P)
IU	20.36	19.79	<b>19.50</b>	19.50	Whisper
JYS	11.79	12.04	11.74	10.69	VICOM-UPM(C2)
NN	<b>9.33</b>	10.20	10.49	9.33	BCN2BRNO(P)
RD	20.60	23.30	20.84	18.18	VICOM-UPM(P)
SYG	<b>10.05</b>	10.24	10.07	10.05	BCN2BRNO(P)
TO	23.98	<b>20.61</b>	24.91	20.61	VICOM-UPM(C1)
YR	29.74	25.33	<b>21.48</b>	21.48	Whisper
TOTAL	<b>14.35</b>	14.78	14.87	14.35	BCN2BRNO(P)

Three shows have been used in previous challenges: CA (Comando Actualidad) and AT (Aquí la Tierra) in 2022 and SyG (Saber y Ganar) in 2018. Table 12 shows the comparison among the 3 last challenges for the three shows and the best system WER over the whole test set. It is noteworthy the significant improvement in WER observed in the three shows and in the whole test set. In the latter case, there is a 10% improvement between the 2020 and 2022 challenges although the organization has tried to rise the difficulty including audio recordings from shows with only raw footage and Spanish Slang. The increase in difficulty is clear if we compare the results obtained in 2020 and 2022 using the Whisper model. Using the VICOM-UPM post-processing of the Whisper output, the WER for the 2020 test set is 12.15% (24.25% relative improvement with regard to the best 2020 system) but for the 2022 test set is 14.87% (22.3% relative increase with respect to the 2020 WER with Whisper).

**Table 12.** WER performance across challenges (2018, 2020 and 2022) on three different shows. The Best System figures are computed on the whole test dataset of each challenge.

Show	2018	2020	2022	WER Improvement
AT	-	13.93%	9.26%	31.1%
CA	-	20.90%	17.71%	15.2%
SyG	14.77%	-	10.05%	31.9%
Best System	16.45%	16.04%	14.35%	10.5%

## 5.2. Speaker Diarization and Identity Assignment Challenge

The results of the Speaker Diarization challenge are presented in Table 13. The table presents the results of the two participating teams and reference results using the best system in the previous 2020 challenge. It is noteworthy the low DER, 18.47%, obtained by IRIT using pyannote.audio. This system gives a 16% of DER over the 2020 dataset, a little

bit higher than the DER, 15.25%, of the best 2020 system. However, the performance of the best 2020 speaker diarization system degrades to 34.29% on the new 2022 test dataset.

**Table 13.** DER (%) performance on the Albayzin 2022 Speaker Diarization Challenge.

System ID	Primary	C1	C2
IV	35.59%	45.92%	48.67%
IRTI	<b>18.47%</b>	19.58%	-
Best System 2020	34.29%	-	-

Table 14 presents diarization results disaggregated by show for the best system in terms of missed speaker error (MiSE), false alarm speaker error (FASE), speaker error (SpE) and diarization error (DER). Again, as in the Speech to Text challenge, the “Agrosfera” show gives the best results with the lowest DER of 8.57%, mainly due to speaker errors. The worst results, as expected, are given by the serial drama shows, with the worst results, 56.47% of DER, for the Yreal show. This show is a thriller with live action and many overlaps of music and speech with regard to the other serial dramas in the test set.

With respect to the Speaker Diarization and Identity Assignment challenge, only the team from Intelligent Voice (IV) presented results. The IV team submitted several systems where the main difference was the VAD and the use of a Deep-Encoder Convolutional Autoencoder Denoiser (DE-CADE) speech enhancement model. The final system submitted by the IV team implemented an SNR classifier to decide whether or not to employ the DE-CADE speech enhancement algorithm on the input audio signals for inference based on the SNR threshold of 0.2, and a final 28.88% AER was reported. Table 15 shows the results of the 4 submissions of the IV team.

**Table 14.** Speaker Diarization 2022 challenge, best system results in terms of the MiSE (missed speaker error), FASE (False Alarm Speaker Error), SpE (Speaker Error) and DER.

Program	#spkrs	MiSE (%)	FASE (%)	SpE (%)	DER (%)
3 × 4	17	7.70	0.77	4.80	13.27
AG	60	0.58	0.25	7.75	8.57
AT	20	4.82	1.02	6.20	12.04
CA	57	7.35	2.19	16.30	25.84
CO	45	2.41	1.05	17.52	20.98
ED	72	7.46	0.64	10.29	18.39
GR	20	11.58	6.37	16.68	34.64
RD	16	15.35	6.69	14.32	36.36
YR	11	26.59	15.18	14.70	56.47
TOTAL		5.86	1.57	11.03	18.47

**Table 15.** Speaker Diarization and Identity Assignment 2022 challenge results in terms of MiSE (missed speaker error), FASE (False Alarm Speaker Error), SpE (Speaker Error) and AER.

	MiSE (%)	FAES (%)	pSE (%)	AER (%)
IV (P)	1.3	229.7	9.5	240.55
IV (C1)	3.7	160.8	20.9	185.42
IV (C2)	8.3	76.5	5.9	90.62
IV (C3)	12.4	15.3	1.2	28.88



### 5.3. Text and Speech Alignment Challenge

#### 5.3.1. Alignment of Re-Speaking Generated Subtitles (TaSAC-ST)

Results for the four systems submitted by GTTS on the development and test sets of TaSAC-ST are shown in Table 16. In all cases, the acoustic models trained on generic databases (Primary and Con-2) perform better than those trained on BP materials (Con-1 and Con-3), suggesting that, despite being three times larger, the BP training dataset does not suitably match the TV broadcasts used in this evaluation. Clearly, using in-domain (RTVE) data to train the acoustic models would lead to better results. On the other hand, the originally submitted systems (Primary and Con-1) show degraded performance on the test set when compared to that obtained on the development set: the APTEM increases by 33.8% and 22.6%, respectively, and the mean of the alignment errors gets multiplied by almost 4, meaning that large alignment errors are being made. The alignment algorithm could be matching some words at the wrong end of relatively long non-speech intervals, thus introducing large alignment errors. This issue (the appearance of relatively long non-speech intervals) would be happening more frequently in the test set than in the development set and would explain the difference in performance. This latter issue gets fixed when using the modified kernel (see details in [27]): the APTEM of Con-2 and Con-3 is almost the same on both (development and test) datasets, while the mean of the alignment errors on the test set gets drastically reduced, from 4.0923 for the Primary system to 0.6053 for Con-2 and from 4.1990 for Con-1 to 0.7186 for Con-3, meaning 85% and 83% relative reductions, respectively.

Table 17 shows the average performance of the best GTTS system (Con-2) on the three TV programs used in the development and test sets of TaSAC-ST, with the aim to discover performance variability depending on the type of TV program. APTEM and the mean of alignment errors are quite similar for AG (Agrosfera) and CO (Corazón), but clearly worse for AT (Aquí la Tierra). Differences are not large in terms of APTEM but remarkable in terms of the mean error, which may indicate that the alignment task could be hard in the case of challenging audio recordings and/or poor subtitles.

**Table 16.** Performance of GTTS primary and contrastive systems on the development (Dev) and test sets of TaSAC-ST. The Average Program Time-Error Metric (APTEM, in seconds) and the global mean of subtitle alignment errors (Mean-Err, in seconds) are shown.

System	Dev		Test	
	APTEM	Mean-Err	APTEM	Mean-Err
Primary	0.2950	1.2665	0.3950	4.0923
Con-1	0.3250	1.1493	0.3986	4.1990
Con-2	0.2950	0.8233	0.2927	0.6053
Con-3	0.3250	0.7950	0.3277	0.7186

**Table 17.** GTTS Con-2 system performance disaggregated per programs: AG (Agrosfera, 10 programs), AT (Aquí la Tierra, 6 programs) and CO (Corazón, 10 programs).

Program	APTEM	Mean-Err
AG	0.2850	0.5250
AT	0.3100	0.9177
CO	0.2910	0.6112

#### 5.3.2. Alignment and Validation of Speech Signals with Partial and Inaccurate Text Transcriptions (TaSAC-BP)

The BUT team did provide development results only for the contrastive system and test results only for the primary system. Performance figures for the baseline and BUT systems are shown in Table 18 (development) and Table 19 (test). The optimal performance

is shown too, along with the optimal word confidence threshold. The search for the optimal performance is done by sorting the accepted words by confidence and then leaving aside words one by one from the bottom of the list (the words with the lowest confidence), measuring the score time in each case and keeping the best one. The confidence threshold in parentheses corresponds to that of the last word accepted, but there could be other words with the same confidence score that were rejected. BUT systems seem to be quite competitive (because they recover most of the available speech in Spanish) and are always better than the baseline. The proposed approach seems to be dealing well with the two languages and matches the reference transcripts (only in Spanish) with the corresponding audio sections. However, since word confidence scores are all set to 1 in BUT submissions, using a confidence threshold to optimize performance does not make sense in this case.

**Table 18.** Performance (time, in seconds) of the baseline and BUT contrastive systems on the TaSAC-BP development set. Optimal performance is shown too (with the optimal threshold in parentheses).

System	Rejected	Accepted	Correct	Wrong	Score
Base	11.97	2410.33	2218.70	191.63	2027.07
Base-opt (0.11)	10.80	2411.50	2219.50	192.00	2027.50
BUT-con	0.00	2458.89	2282.26	176.63	2105.63
BUT-con-opt (1.00)	109.21	2349.68	2252.09	97.59	2154.50

#### 5.4. Search on Speech Challenge

The results of the search on speech challenge are presented in Table 20 for the RTVE2022DB test dataset, which was the only dataset that received submissions. They show that the Multistream CNN+rescoring system performs better than the Multistream CNN system. A paired *t*-test shows that this better performance is significant ( $p < 0.02$ ). This indicates that the rescoring approach presented in the Multistream CNN+rescoring system helps for term detection. The small performance gap between MTWV and ATWV results clearly indicates that the detection threshold is well-calibrated. When comparing the best evaluation result (ATWV = 0.6694) with the best system submitted to the 2020 SoS challenge (ATWV = 0.2123), the performance is much better. This is due to the more robust ASR system constructed for this year's evaluation.

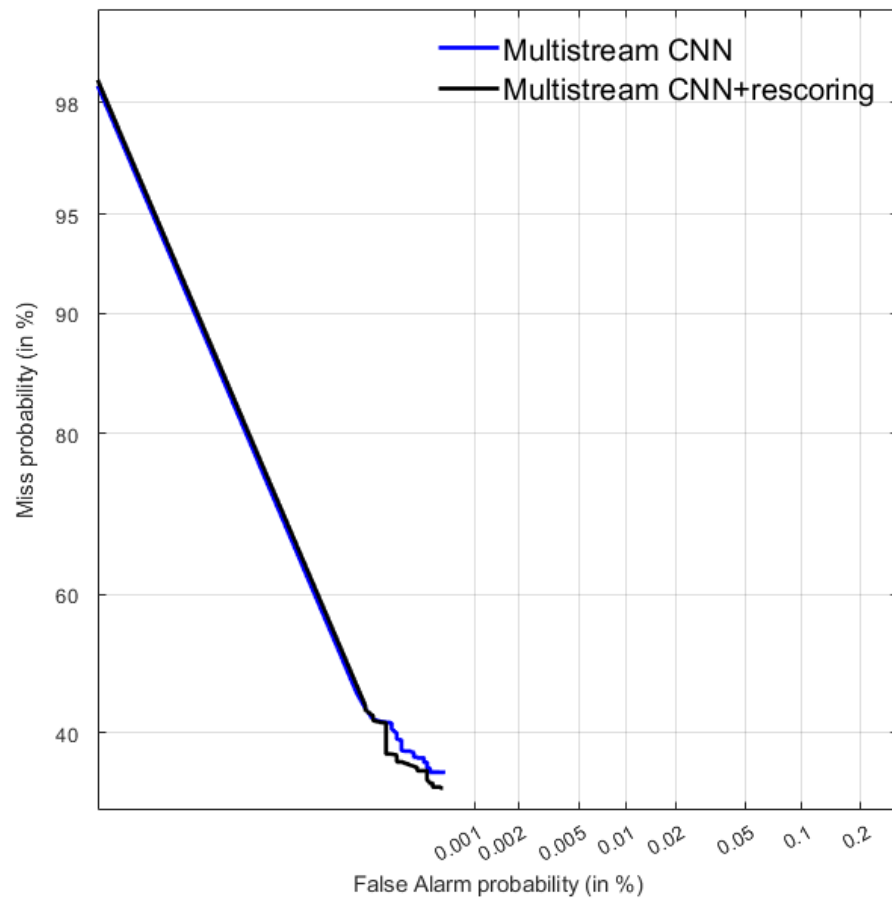
**Table 19.** Performance (time, in seconds) of the baseline and BUT primary systems on the TaSAC-BP test set. Optimal performance is shown too (with the optimal threshold in parentheses).

System	Rejected	Accepted	Correct	Wrong	Score
Base	4.63	1794.82	1628.96	165.86	1463.10
Base-opt (0.25)	5.85	1793.60	1628.51	165.09	1463.42
BUT-pri	0.00	1820.98	1688.50	132.48	1556.02
BUT-pri-opt (1.00)	72.13	1748.85	1668.58	80.27	1588.31

**Table 20.** Search on speech challenge results for the RTVE2022DB test dataset.

System ID	MTWV	ATWV	p(FA)	p(Miss)
Multistream CNN	0.6496	0.6483	0.00000	0.346
Multistream CNN + rescoring	0.6696	0.6694	0.00001	0.325

The DET curves of the submitted systems are presented in Figure 1. They show that the Multistream CNN+rescoring system does generally perform the same as the Multistream CNN system for high miss ratios and does outperform the Multistream CNN system for low miss ratios.



**Figure 1.** Search on speech challenge: DET curves. DET curves for the systems submitted to the search on speech challenge for the RTVE2022DB test dataset.

## 6. Conclusions

The Albayzin evaluation campaigns comprise a unique and closed framework to measure the progress of speech technologies in Iberian languages (especially Spanish). This paper provides a comprehensive overview of the most recent Albayzin evaluation campaign: the IberSpeech-RTVE 2022 Challenges, which featured four different evaluations focused on TV broadcast content: speech-to-text transcription, speaker diarization and identity assignment, text and speech alignment, and search on speech. This is the first time that the text and speech alignment task is addressed in the Albayzin evaluation campaigns. Since its first release in 2018, the RTVE database provided for system development has increased in size up to 1000 h, with half of the material human transcribed, 96 h labelled with speaker turns and 58 h annotated with identity labels from a closed set of around 200 characters. Besides RTVE broadcast data, conference talks and panels (MAVIR) and parliamentary sessions (SPARL22, Basque Parliament dataset) have been also used in some of the challenges, to expand the range of domains and conditions.

In the Speech-To-Text task, four teams participated with 13 different system submissions. The evaluation was carried out over 21 different shows with a total of 54 h and covering a broad range of acoustic conditions. The best results in terms of the lowest WER were given by a system composed of the fusion of 5 models with a WER of 14.35%. The best single system gave a WER of 14.78%. It is remarkable that the zero-shot system, the Whisper Large model, released in September 2022, when using proper post-processing of the output, obtained a WER of 14.87%. It is also noteworthy that the 2022 test dataset is more difficult than the previous one by a 22.3%, according to the results obtained using Whisper in both datasets.

With regard to the speaker diarization and identity assignment challenge, only two teams participated in the speaker diarization task and only one of them submitted systems to the identity assignment task. The organization provided a baseline system based on the best 2020 system. The evaluation was carried out over 9 different shows with a total of 25 h and covered a broad range of conditions: acoustic background, number of speakers, and amount of overlapping speech. The best system obtained 18.47% DER, which meant a big gap in terms of DER when compared to the other submissions, including the baseline system.

There were only two participants in the text and speech alignment challenge. In the first task, focused on re-speaking generated subtitles, the best submitted system obtained an average median error per program (APTEM) of 0.29 s, and a global mean error of 0.61 s, which are still too high for many applications. The system was trained on out-of-domain (non-RTVE) data so better results could be expected when using in-domain training data. Also, it was found that some programs could be more challenging than others, with almost twice global mean error. In the second task, focused on retrieving training materials from loosely transcribed audio involving two languages (Basque and Spanish), the submitted system leveraged state-of-the-art ASR technology to perform unrestricted text and speech alignments, obtaining very competitive scores and successfully matching most of the reference transcripts (only in Spanish) with the corresponding audio sections.

In the search on speech challenge, which employed a subset of the test dataset employed in the speech-to-text transcription task, a single participant submitted two different systems. The best system obtained an ATWV of 0.6694, which shows that there is still ample room for improvement in this task.

In summary, the results are showing an improvement in the performance of the speech technologies assessed in the four challenges comparing with previous challenges. Promising results in both speech-to-text transcription and speaker diarization tasks were obtained in some TV shows belonging to genres such as news, interviews, thematic magazines or daily magazines. However, there is still room for improvement when dealing with genres such as serial drama, reality shows and game shows. For text and speech alignment, more datasets are needed in order to cover all the problems associated to re-speaking where the re-speakers often end up paraphrasing, and more in-the-wild materials involving a wide range of acoustic conditions and transcript qualities should be used to assess the performance of text and speech alignment methods for leveraging loosely or partially transcribed audio resources as training data for ASR. With regard to the search on speech tasks (STD and QbE-STD), more research is needed to address the increased complexity posed by out-of-vocabulary terms and/or multi-word terms when compared to in-vocabulary and single-word terms.

We are now actively preparing a new edition of the Albayzin evaluations, to be held along with the next IberSpeech conference in 2024. An extension of the RTVE database with new challenging audiovisual material will be released by April 2024, which will hopefully help to assess new developments in speech and language technologies.

**Author Contributions:** Conceptualization, E.L., V.B., L.J.R.-F. and J.T.; methodology, E.L., L.J.R.-F. and J.T.; software, E.L., L.J.R.-F., A.O., A.M., V.B., C.P., A.d.P., M.P., A.V., G.B., A.Á., H.A. and D.T.-T.; validation, E.L., L.J.R.-F., J.T. and D.T.-T.; formal analysis, E. L., L.J.R.-F., J.T. and D.T.-T.; investigation, E.L., L.J.R.-F., J.T. and D.T.-T.; resources, all authors; data curation, E.L., L.J.R.-F., J.T. and D.T.-T.; writing—original draft preparation, E.L., L.J.R.-F. and J.T.; writing—review and editing, all authors; visualization, all authors; supervision, E.L., L.J.R.-F. and J.T.; project administration, E.L., V.B., C.P. and A.d.P.; funding acquisition, E.L., V.B., C.P. and A.d.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was partially supported by Radio Televisión Española through the RTVE Chair at the University of Zaragoza, and Red Temática en Tecnologías del Habla (RED2022-134270-T), funded by AEI (Ministerio de Ciencia e Innovación); It was also partially funded by the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie Grant 101007666; in part by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU"/

PRTR under Grants PDC2021-120846C41 PID2021-126061OB-C44, and in part by the Government of Aragon (Grant Group T3623R); it was also partially funded by the Spanish Ministry of Science and Innovation (OPEN-SPEECH project, PID2019-106424RB-I00) and by the Basque Government under the general support program to research groups (IT-1704-22), and by projects RTI2018-098091-B-I00 and PID2021-125943OB-I00 (Spanish Ministry of Science and Innovation and ERDF) as well.

**Institutional Review Board Statement:** Not applicable

**Informed Consent Statement:** Not applicable

**Data Availability Statement:** The RTVE database is freely available subject to the terms of a license agreement with RTVE (<http://catedrartve.unizar.es/rtvedatabase.html> (accessed on 24 July 2023)). Requirements for downloading the MAVIR database can be found in <http://cartago.llf.uam.es/mavir/index.pl?m=descargas> (accessed on 24 July 2023). For details on SPARL22 database access, please contact Javier Tejedor ([javier.tejedornoguerales@ceu.es](mailto:javier.tejedornoguerales@ceu.es)). To access the Basque Parliament datasets, the ground-truth files and the evaluation script of the TaSAC-BP evaluation, please contact Luis Javier Rodriguez-Fuentes ([luisjavier.rodriguez@ehu.eus](mailto:luisjavier.rodriguez@ehu.eus)).

**Acknowledgments:** We gratefully acknowledge the support of the IberSpeech 2022 organizers and every one of the participants for their invaluable contribution to the Albayzin evaluations.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## References

1. Garofolo, J.; Fiscus, J.; Fisher, W. Design and preparation of the 1996 HUB-4 broadcast news benchmark test corpora. In *Proceedings of the DARPA Speech Recognition Workshop*; Morgan Kaufmann Publishers: Burlington, MA, USA, 1997.
2. Graff, D. An overview of Broadcast News corpora. *Speech Commun.* **2002**, *37*, 15–26. [https://doi.org/10.1016/S0167-6393\(01\)00057-7](https://doi.org/10.1016/S0167-6393(01)00057-7).
3. Bell, P.; Gales, M.J.F.; Hain, T.; Kilgour, J.; Lanchantin, P.; Liu, X.; McParland, A.; Renals, S.; Saz, O.; Wester, M.; et al. The MGB challenge: Evaluating multi-genre broadcast media recognition. In *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, USA, 13–17 December 2015; pp. 687–693. <https://doi.org/10.1109/ASRU.2015.7404863>.
4. NIST. *NIST Open Keyword Search 2013 Evaluation (OpenKWS13)*, 1st ed.; National Institute of Standards and Technology (NIST): Washington, DC, USA, 2013.
5. NIST. *NIST Open Keyword Search 2014 Evaluation (OpenKWS14)*, 1st ed.; National Institute of Standards and Technology (NIST): Washington, DC, USA, 2014.
6. NIST. *NIST Open Keyword Search 2015 Evaluation (OpenKWS15)*, 1st ed.; National Institute of Standards and Technology (NIST): Washington, DC, USA, 2015.
7. NIST. *NIST Open Keyword Search 2016 Evaluation (OpenKWS16)*, 1st ed.; National Institute of Standards and Technology (NIST): Washington, DC, USA, 2016.
8. NIST. *2017 Pilot Open Speech Analytic Technologies Evaluation (2017 NIST Pilot OpenSAT)*, 1st ed.; National Institute of Standards and Technology (NIST): Washington, DC, USA, 2019.
9. NIST. *NIST Open Speech Analytic Technologies 2019 Evaluation Plan (OpenSAT19)*, 1st ed.; National Institute of Standards and Technology (NIST): Washington, DC, USA, 2019.
10. NIST. *NIST Open Speech Analytic Technologies 2020 Evaluation Plan (OpenSAT20)*, 1st ed.; National Institute of Standards and Technology (NIST): Washington, DC, USA, 2020.
11. 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. <https://doi.org/10.3390/app9245412>.
12. Lleida, E.; Ortega, A.; Miguel, A.; Bazán-Gil, V.; Pérez, C.; Gómez, M.; de Prada, A. IberSpeech 2020 Evaluation Results. Available online: <http://catedrartve.unizar.es/albayzin2020results.html> (accessed on 22 June 2023).
13. Lleida, E.; Ortega, A.; Miguel, A.; Bazán, V.; Pérez, C.; Gómez, M.; de Prada, A. RTVE2018 Database Description. Available online: <http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf> (accessed on 22 June 2023).
14. Lleida, E.; Ortega, A.; Miguel, A.; Bazán-Gil, V.; Pérez, C.; Gómez, M.; de Prada, A. RTVE2020 Database Description. Available online: <http://catedrartve.unizar.es/reto2020/RTVE2020DB.pdf> (accessed on 22 June 2023).
15. Zelenák, M.; Schulz, H.; Hernando, J. Albayzin 2010 Evaluation campaign: Speaker diarization. In *Proceedings of the Jornadas en Tecnología del Habla and Iberian SLTech Workshop*, Vigo, Spain, 10–12 November 2010; p. 301.
16. Zelenák, M.; Schulz, H.; Hernando, J. Speaker diarization of broadcast news in Albayzin 2010 evaluation campaign. *EURASIP J. Audio Speech Music. Process.* **2012**, *2012*, 19. <https://doi.org/10.1186/1687-4722-2012-19>.

17. Fiscus, J.G.; Ajot, J.G.; Garofolo, J.S.; Doddington, G. Results of the 2006 Spoken Term Detection Evaluation. In Proceedings of the ACM SIGIR, Amsterdam, The Netherlands, 23–27 July 2007; pp. 1–4.
18. Metz, F.; Anguera, X.; Barnard, E.; Davel, M.; Gravier, G. Language Independent Search in Mediaeval’s Spoken Web Search Task. *Comput. Speech Lang.* **2014**, *28*, 1066–1082. <https://doi.org/10.1016/j.csl.2013.12.004>.
19. Martin, A.; Doddington, G.; Kamm, T.; Ordowski, M.; Przybocki, M. The DET Curve In Assessment Of Detection Task Performance. In Proceedings of the Eurospeech, Rhodes, Greece, 22–25 September 1997; pp. 1895–1898.
20. NIST. *Evaluation Toolkit (STDEval) Software*; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 1996.
21. Kocour, M.; Umesh, J.; Karafiat, M.; Švec, J.; López, F.; Luque, J.; Beneš, K.; Diez, M.; Szoke, I.; Veselý, K.; et al. BCN2BRNO: ASR System Fusion for Albayzin 2022 Speech to Text Challenge. In Proceedings of the IberSPEECH, Granada, Spain, 14–16 November 2022; pp. 276–280. <https://doi.org/10.21437/IberSPEECH.2022-56>.
22. Miguel, A.; Ortega, A.; Lleida, E. ViVoLAB System Description for the S2TC IberSPEECH-RTVE 2022 Challenge. Available online: [http://catedrartve.unizar.es/reto2022/83-ViVoLAB\\_System\\_Description\\_for\\_S2TC\\_IberSPEECH\\_RTVE\\_2022\\_challenge.pdf](http://catedrartve.unizar.es/reto2022/83-ViVoLAB_System_Description_for_S2TC_IberSPEECH_RTVE_2022_challenge.pdf) (accessed on 23 June 2023).
23. López, F.; Luque, J. TID Spanish ASR system for the Albayzin 2022 Speech-to-Text Transcription Challenge. In Proceedings of the IberSPEECH, Granada, Spain, 14–16 November 2022; pp. 271–275. <https://doi.org/10.21437/IberSPEECH.2022-55>.
24. Arzelus, H.; Torres, I.G.; Martín-Doñas, J.M.; González-Docasal, A.; Alvarez, A. The Vicomtech-UPM Speech Transcription Systems for the Albayzín-RTVE 2022 Speech to Text Transcription Challenge. In Proceedings of the IberSPEECH, Granada, Spain, 14–16 November 2022; pp. 266–270. <https://doi.org/10.21437/IberSPEECH.2022-54>.
25. Bredin, H. Pyannote.Audio 2.1 Speaker Diarization Pipeline: Principle, Benchmark, and Recipe. Available online: [https://catedrartve.unizar.es/reto2022/PYA\\_report.pdf](https://catedrartve.unizar.es/reto2022/PYA_report.pdf) (accessed on 22 June 2023).
26. Shrestha, R.; Glackin, C.; Wall, J.; Moniri, M.; Canning, N. Intelligent Voice Speaker Recognition and Diarization System for IberSpeech 2022 Albayzin Evaluations Speaker Diarization and Identity Assignment Challenge. Available online: [https://catedrartve.unizar.es/reto2022/82-Albayzin\\_IV\\_paper\\_final.pdf](https://catedrartve.unizar.es/reto2022/82-Albayzin_IV_paper_final.pdf) (accessed on 22 June 2023).
27. Bordel, G.; Rodríguez-Fuentes, L.J.; Peñagarikano, M.; Varona, A. GTTS Systems for the Albayzin 2022 Speech and Text Alignment Challenge. In Proceedings of the IberSPEECH, Granada, Spain, 14–16 November 2022; pp. 285–289. <https://doi.org/10.21437/IberSPEECH.2022-58>.
28. Collobert, R.; Puhersch, C.; Synnaeve, G. Wav2Letter: An End-to-End ConvNet-based Speech Recognition System. *CoRR* **2016**, abs/1609.03193.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.