applied sciences

MDPI

Article

Albayzin 2018 Evaluation: The IberSpeech-RTVE Challenge on Speech Technologies for Spanish Broadcast Media

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

[1] Vivolab, Aragon Institute for Engineering Research (I3A), University of Zaragoza, 50018 Zaragoza, Spain; ortega@unizar.es (A.O.); amiguel@unizar.es (A.M.)

[2] Corporación Radiotelevisión Española, 28223 Madrid, Spain; virginia.bazan@rtve.es (V.B.-G.); carmen.perez.cernuda@rtve.es (C.P.); manuel.gomez@rtve.es (M.G.); alberto.deprada@rtve.es (A.d.P.)

* Correspondence: lleida@unizar.es

Received: 18 October 2019; Accepted: 3 December 2019; Published: 11 December 2019

Abstract: The IberSpeech-RTVE Challenge presented at IberSpeech 2018 is a new Albayzin evaluation series supported by the Spanish Thematic Network on Speech Technologies (Red Temática en Tecnologías del Habla (RTTH)). That series was focused on speech-to-text transcription, speaker diarization, and multimodal diarization of television programs. For this purpose, the Corporacion Radio Television Española (RTVE), the main public service broadcaster in Spain, and the RTVE Chair at the University of Zaragoza made more than 500 h of broadcast content and subtitles available for scientists. The dataset included about 20 programs of different kinds and topics produced and broadcast by RTVE between 2015 and 2018. The programs presented different challenges from the point of view of speech technologies such as: the diversity of Spanish accents, overlapping speech, spontaneous speech, acoustic variability, background noise, or specific vocabulary. This paper describes the database and the evaluation process and summarizes the results obtained.

Keywords: IberSpeech Challenge; RTVE2018 database; Albayzin evaluation; speech-to-text transcription; speaker diarization; multimodal diarization

1. Introduction

Albayzin is a series of technological evaluations open to the scientific community in order to propose challenges and datasets to work with in different fields of the broad area of speech technologies. Organized since 2006 and supported by the Spanish Thematic Network on Speech Technologies (RTTH) (Red Temática en Tecnologías del Habla (RTTH) http://www.rthabla.es), in 2018, the broadcast media area was addressed. Jointly with Radio Televisión Española, RTVE (Radiotelevisión Española (RTVE): http://www.rtve.es), the Spanish Public Broadcast Corporation, Vivolab (Vivolab http://vivolab.unizar.es), the speech research group at the University of Zaragoza, proposed a set of technological evaluations in the areas of speaker diarization and speech and face recognition. These evaluations were also supported by the RTVE Chair at the University of Zaragoza (Cátedra RTVE de la Universidad de Zaragoza: http://catedrartve.unizar.es). The dataset provided to participants included more than 500 h of broadcast content, spanning a broad range of genres. In addition, part of the media content was provided with subtitles, human revised transcriptions, and speaker labels.

Since 1996, when DARPA [1] presented the HUB-4 broadcast news corpora, several evaluations for broadcast speech related tasks have been organized, most of them in English [2]. Campaigns have been also carried out in other languages such as the ESTER evaluations in French [3,4], the Technology and Corpora for Speech to Speech Translation (TC-STAR) evaluation in Mandarin [5], the National Institute of Standards and Technology (NIST) Rich Transcription evaluations in 2003 and 2004 with data in English, Mandarin, and Arabic, the Albayzin 2010 evaluation campaign in Catalan [6,7], the Albayzin 2012, 2014, and 2016 in Spanish [8–11], and more recently, the Multi-Genre Broadcast (MGB) Challenge with data in English and Arabic 2 [12–14]. In other areas apart from broadcast speech, several evaluation campaigns have been proposed such as the ones organized in the scope of the Zero Resource Speech Challenge [15,16], the TC-STAR evaluation on recordings of the European Parliament's sessions in English and Spanish [5], or the MediaEval evaluation of multimodal search and hyperlinking [17].

As a way to measure the performance of different techniques and approaches, in this 2018 edition, the IberSpeech-RTVE Challenge Evaluation campaign was proposed in three different conditions: speech-to-text transcription (STT), speaker diarization (SD), and multimodal diarization (MD). Twenty-two teams registered to the challenge, and eighteen submitted systems in at least one of the three proposed tasks. In this paper, we describe the challenge and the data provided by the organization to the participants. We also provide a description of the systems presented to the evaluation, their results, and a set of conclusions that can be drawn from this evaluation campaign.

This paper is organized as follows. In Section 2, the RTVE2018 database is presented. Section 3 describes the three evaluation tasks, speech-to-text transcription, speaker diarization, and multimodal diarization. Section 4 provides a brief description of the main features of the submitted systems. Section 5 presents results, and Section 6 gives conclusions.

2. IberSpeech-RTVE 2018 Evaluation Data

The RTVE2018 database is a collection of TV shows that belong to diverse genres and broadcast by the public Spanish Television (RTVE) from 2015 to 2018. Table 1 presents the titles, duration, and content of the shows included in the RTVE2018 database. The database is composed of 569 h and 22 min of audio. About 460 h are provided with the subtitles, and about 109 h have been human transcribed. We would like to highlight that in most of the cases, subtitles do not contain verbatim transcriptions of the audio since most of them were generated by a re-speaking procedure (The re-speaker re-utters everything that is being said to a speech-to-text transcription system. Most of the time, the re-speaker summarizes what is being said.). The corpus is divided into four partitions, a training one, two development partitions (dev1 and dev2) and finally, a test partition. Additionally, the corpus includes a set of text files extracted from all the subtitles broadcast by the RTVE 24H Channel during 2017.

The training partition consists of all the audio files without human revised transcriptions, which means that only subtitles are available. The training partition can be used for any evaluation task. For development, two partitions are defined. Partition dev1 contains about 53 h of audio and their corresponding human revised transcriptions. The dev1 partition can be used for either development or training of the speech-to-text systems. Partition dev2 contains about 15 h of audio, human revised transcriptions, speaker changes, and their corresponding speaker labels. Additionally, dev2 contains a 2 h show annotated for multimodal diarization (face and speaker) and enrollment files (pictures, videos, and audio) needed for speaker and face identification. Table 2 shows detailed information about the shows included in the development partitions.

The RTVE2018 database includes a test partition with all the files needed to evaluate systems for speech-to-text and speaker and multimodal diarization. Table 3 presents the content of the test partition. The test set covers diverse genres from broadcast news, live magazines, quiz shows, to documentary series with a diversity of acoustic scenarios. Additionally, the test partition contains the enrollment files for the multimodal diarization challenge. It consists of 10 pictures and a 20 second video of the 39 characters to be identified.

Further detailed information about the RTVE2018 database content and formats can be found in the RTVE2018 database description report (http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf).

The RTVE2018 database is freely available subject to the terms of a license agreement with RTVE (http://catedrartve.unizar.es/rtvedatabase.html).

Table 1. Information about the shows included in the RTVE2018 database.

Show	Duration	Show Content
20H	41:35:50	News of the day.
Agrosfera	37:34:32	Agrosfera wants to bring the news of the countryside and the sea to farmers, ranchers, fishermen, and rural inhabitants. The program also aims to bring this rural world closer to those who do not inhabit it, but who do enjoy it.
Al filo de lo Imposible	11:09:57	This show broadcasts documentaries about mountaineering, climbing, and other outdoor risky sports. It is a documentary series in which emotion, adventure, sports, and risk predominate.
Arranca en Verde	05:38:05	Contest dedicated to road safety. In it, viewers are presented with questions related to road safety in order to disseminate in a pleasant way the rules of the road and thus raise awareness about civic driving and respect for the environment.
Asuntos Públicos	69:38:00	All the analysis of the news of the day and the live broadcast of the most outstanding information events.
Comando Actualidad	17:03:41	A show that presents a current topic through the choral gaze of several street reporters. Four journalists who travel to the place where the news occurs show them as they are and bring their personal perspective to the subject.
Dicho y Hecho	10:06:00	Game show in which a group of 6 comedians and celebrities compete against each other through hilarious challenges.
España en Comunidad	13:02:59	Show that offers in-depth reports and current information about the different Spanish autonomous communities. It is made by the territorial and production centers of RTVE.
La Mañana	227:47:00	Live show, with a varied offer of content for the whole family and with the clear vocation of public service.
La Tarde en 24H Economia	04:10:54	Program about the economy.
La Tarde en 24H Tertulia	26:42:00	Talk show of political and economic news (4 to 5 people).
La Tarde en 24H Entrevista	04:54:03	In-depth interviews with personalities from different fields.
La Tarde en 24H el Tiempo	02:20:12	Weather information of Spain, Europe, and America.
Latinoamérica en 24H	16:19:00	Analysis and information show focused on Ibero-America, in collaboration with the information services of the international area and the network of correspondents of RTVE.
Millennium	19:08:35	Debate show of ideas that pretends to be useful to the spectators of today, accompanying them in the analysis of everyday events.
Saber y Ganar	29:00:10	Daily contest presented that aims to disseminate culture in an entertaining way. Three contestants demonstrate their knowledge and mental agility, through a set of general questions.
La Noche en 24H	33:11:06	Talk show with the best analysts to understand what has happened throughout the day. It contains interviews with some of the protagonists of the day.
Total duration	**569:22:04**	

Table 2. Development (dev) dataset partition with shows and duration. S2T, speech-to-text; SD, speaker diarization; MD, multimodal diarization.

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

Table 3. Test dataset partition with shows and duration.

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

3. IberSpeech-RTVE 2018 Evaluation Tasks

This section presents a brief summary of the three evaluation tasks. A more detailed description of the evaluation plans can be found on the Interspeech2018 web page (http://iberspeech2018.talp.cat/index.php/albayzin-evaluation-challenges/) or the Cátedra RTVE-UZ web page http://catedrartve.unizar.es/reto2018/evaluations2018.html.

3.1. Speech-to-Text Challenge

3.1.1. Challenge Description and Databases

The speech-to-text transcription evaluation consisted of automatically transcribing different types of TV shows. The main objective was to evaluate the state-of-the-art in automatic speech recognition (ASR) for the Spanish language in the broadcast sector.

Training and Development Data

The training partition consisted of all the audio files without human revised transcriptions, which means that only subtitles were available. The training partition contained up to 460 h of audio, half of them corresponding to a live show (“La Mañana”). Participants were free to use this audio as they considered appropriate.

For development, two partitions were defined. Partition dev1 contains about 53 h of audio and their corresponding human revised transcriptions. Partition dev2 contains about 15 h of audio and the corresponding human revised transcriptions and speaker change timestamps. For this challenge, both partitions could be used for either development or training.

Training Conditions

The speech-to-text systems could be evaluated over a closed-set or open-set training condition.

- Closed-set condition: The closed-set condition limited the system training to use the training and development datasets of the RTVE2018 database. The use of pretrained models on data other than RTVE2018 was not allowed in this condition. Participants could use any external phonetic transcription dictionary.

- Open-set condition: The open-set training condition removed the limitations of the closed-set condition. Participants were free to use the RTVE2018 training and development set or any other data to train their systems provided that these data were fully documented in the system's description paper.

Each participant team should submit at least a primary system in one condition, open-set or closed-set , but they could also submit up to two contrastive systems.

Evaluation Data

The evaluation data contained a set of eight different TV shows covering a variety of scenarios with a total of 39:07 h of audio (see Table 3). The selected shows were different from those included in the development partition with human revised transcriptions. Table 4 shows the main characteristics of the selected shows.

Table 4. S2T test dataset characteristics.

Show	Acronym	# of Shows	Duration	Audio Features
Al filo de lo Imposible	AFI	9	4:10:03	Poor quality audio in some outdoor shots. Few speakers. Exterior shots.
Arranca en Verde	AV	2	1:00:30	Good audio quality in general. Most of the time, 2 speakers in a car.
Dicho y Hecho	DH	1	1:48:00	Much speech overlap and speech inflections. About 8 speakers, most of them comedians. Studio and exterior shots.
España en Comunidad	EC	22	8:09:32	Good audio quality in general. Diversity of speakers. Studio and exterior shots.
La Mañana	LM	4	8:05:00	Much speech overlap, speech inflections, and live audio. Studio and exterior shots.
La Tarde en 24H (Tertulia)	LT24HTer	9	8:52:20	Good audio quality, overlapped speech on rare occasions, up to 5 speakers. Television studio.
Latinoamérica en 24H	LA24H	8	4:06:57	Good audio quality. Many speakers with a Spanish Latin American accent. Studio and exterior shots.
Saber y Ganar	SG	4	2:54:53	Good audio quality. Up to 6 speakers per show. Television studio.
		59	**39:07:15**	

3.1.2. Performance Measurement

The STT system output was evaluated with different metrics, but they were ranked by the word error rate. All the participants had to provide as output for evaluation a free-form text with no page, paragraphs, sentence, or speaker breaks using the UTF-8 charset http://www.utf-8.com/ per test file. The text might include punctuation marks to be evaluated with an alternative metric.

Primary Metric

The word error rate (WER) was the primary metric for the STT task. The text was normalized removing all the punctuation marks; numbers were written with letters; and text was lower-cased. The WER is defined as:

$$WER = \frac{S + D + I}{N_r} \tag{1}$$

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.

Alternative Metrics

In addition to the primary metric, other alternative metrics were computed, but not taken into account for the challenge ranking.

Punctuation marks evaluation (PWER): The WER was computed with the punctuation marks given by the STT system. Periods and commas were processed as words.

Text normalized word error rate (TNWER): Text normalization techniques such as stop-word removal and lemmatization were applied to the STT output. In this sense, common errors such as verbal conjugations, gender or number substitutions, articles, determiners, and quantifiers deletion/insertions had no impact on the performance evaluation metric. The same text normalization was applied to both the reference and automatic transcriptions before proceeding to calculate WER. The Freeling http://nlp.lsi.upc.edu/freeling/ lemmatizer was used.

3.2. Speaker Diarization Challenge

3.2.1. Challenge Description and Databases

The speaker diarization challenge consisted of segmenting broadcast audio documents according to different speakers and linking those segments that originated from the same speaker. No a priori knowledge was provided about the number or the identity of the speakers participating in the audio to be analyzed. The diarization error rate (DER) was used as the scoring metric as defined in the Rich Transcription (RT) evaluations organized by NIST. The open-set and closed-set training conditions were proposed in the challenge. Participants could submit systems in one or both conditions in an independent way.

Databases

RTVE2018 database: The RTVE2018 training and development partitions might be used for any purpose including system development or training. In particular, the dev2 partition includes around sixteen hours with diarization and reference speech segmentation corresponding to two different debate shows. There are four episodes (7:26 h) of "La noche en 24H" http://www.rtve.es/alacarta/videos/la-noche-en-24-horas/, where a group of political analysts comments about the news of the day, and eight episodes (7:42 h) of "Millennium" http://www.rtve.es/alacarta/videos/millennium/, where a group of experts debates a current issue.

Aragón Radio database: The database donated by the *Corporación Aragonesa de Radio y Televisión* http://www.cartv.es/ (CARTV) consisted of around twenty hours of Aragón Radio broadcast. This dataset contains around 85% of speech, 62% of music, and 30% of noise in such a way that 35% of the audio contains music along with speech, 13% is noise along with speech, and 22% is speech alone.

3/24 TV channel database: The Catalan broadcast news database from the 3/24 TV channel http://www.ccma.cat/324/ proposed for the 2010 Albayzin Audio Segmentation Evaluation [7,18] was recorded by the Language and Speech Technologies and Applications Center (TALP) from the

Universitat Politècnica de Catalunya (UPC) in 2009 under the Tecnoparla project http://rua.ua.es/dspace/handle/10045/8626 funded by the Generalitat de Catalunya. The *Corporació Catalana de Mitjans Audiovisuals* http://www.ccma.cat (CCMA), owner of the multimedia content, allowed its use for technology research and development. The database consists of around 87 h of recordings in which speech can be found in 92% of the segments, music is present 20% of the time, and noise in the background 40%. Another class called *others* was defined, which can be found 3% of the time. Regarding the overlapped classes, 40% of the time, speech can be found along with noise and 15% of the time, speech along with music.

Training Conditions

Two training conditions, closed-set and open-set, were proposed:

- Closed-set condition: The closed-set condition limited the use of data to the set of audio of the three partitions distributed in the challenge.
- Open-set condition: The open-set condition eliminated the limitations of the closed-set condition. Participants were free to use any dataset, as long as they were publicly accessible for all (not necessarily free).

Evaluation Data

The evaluation of the diarization systems was done exclusively using the evaluation partition of the RTVE2018 database. This partition consisted of 22:45 h of various programs (see Table 3), 22 episodes of "España en comunidad", which corresponds to 35.47% of the total audio, 8 episodes of "Latinoamerica en 24H" with 17.83%, 1 episode of "La Mañana", which represents 7.19% of the total, and 9 episodes of "La Tarde en 24H", which represents 39.50% of the total audio. No a priori knowledge was provided about the number or the identity of speakers participating in the audio to be analyzed.

3.2.2. Diarization Scoring

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

Given the dataset to evaluate Ω, each document was 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 was defined as:

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

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.

$$DER = \frac{\sum\limits_{n\in\Omega} E(n)}{\sum\limits_{n\in\Omega}\left(T(n)N_{ref}(n)\right)} \tag{3}$$

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. This error can occur in segments where the number of system speakers is greater than the number of reference speakers, but also in segments where the number of system speakers is lower than the number of reference speakers whenever the number of system speakers and the number of reference speakers are greater than zero.
- Missed speech time: The missed speech time refers to the amount of time that speech is present, but not labeled by the diarization system in segments where the number of system speakers is lower than the number of reference speakers.
- False alarm time: The false alarm time is the amount of time that a speaker has been labeled by the diarization system, but is not present in segments where the number of system speakers is greater than the number of reference speakers.

Consecutive speech segments of audio labeled with the same speaker identification tag and separated by a non-speech segment less than 2 s long were merged and considered a single segment. A region of 0.25 s around each segment boundary, usually known as the forgiveness collar, was considered. These regions were 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.

3.3. Multimodal Diarization Challenge

3.3.1. Challenge Description and Databases

The multimodal diarization evaluation consisted of segmenting broadcast audiovisual documents according to a closed set of different speakers and faces and linking those segments that originated from the same speaker and face. For this evaluation, a list of characters to recognize was given. The rest of the characters on the audiovisual document were discarded for evaluation purposes. System outputs should give for each segment who was speaking and who was/were in the image from the list of characters. For each character, a set of face pictures and a short audiovisual document were given.

The goal of this challenge was to start a new series of Albayzin evaluations based on multimodal information. In this edition, we focused on face and speaker diarization. We wanted to evaluate the use of audiovisual information for speaker and face diarization. We encouraged participants to use both speaker and face information jointly for diarization, although we accepted systems that used visual and audio information separately.

Development and Evaluation Data

- For development, the dev2 partition contained a two-hour show "La noche en 24H" labeled with speaker and face timestamps. Enrollment files for the main characters were also provided. Enrollment files consisted of pictures and short videos with the character speaking. Additionally, the dev2 partition contained around 14 h of speaker diarization timestamps. No restrictions were placed on the use of any data outside the RTVE2018.
- For the evaluation, three television programs were distributed, one from "La Mañana" and two from "La Tarde en 24H Tertulia", which totaled four hours. For enrollment, photos (10) and video (20 s) of the 39 characters to be labeled were provided.

3.3.2. Performance Scoring

The multimodal diarization performance scoring evaluated the accuracy of indexing a TV show in terms of the amount of people speaking and present in the image. To measure the performance of the proposed systems, DER was computed as the fraction of speaker or face time that was not correctly attributed to that specific character. This score was computed over the entire file to be processed; including regions where more than one character was present (overlap regions).

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

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

As in the speaker diarization task, consecutive speech segments of the same speaker separated by a non-speech segment of less than 2 s long were merged into a single segment, and a forgiveness collar of 0.25 s around each speaker or face segment boundary was considered.

The primary metric to rank systems was the average of the face and speaker diarization errors:

$$DER_{total} = 0.5DER_{spk} + 0.5DER_{face} \tag{4}$$

4. Submitted Systems

4.1. Speech-to-Text Challenge

A total of 20 different systems from seven participating teams was submitted. All of them presented results in the open-set condition, and three of them also presented results in the closed-set condition.

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 of STT systems.

4.1.1. Open-Set Condition Systems

- G1-GTM-UVIGO [19]. Multimedia Technologies Group, Universidad de Vigo, Spain.
  G1-GTM-UVIGO submitted two systems using as the recognition engine the Kaldi toolkit http://kaldi-asr.org/. Primary and contrastive systems differed in the language model (LM) used in the rescoring stage. The primary system used the four-gram LM, and the contrastive system used the recurrent neural network language modeling toolkit (RNNLM) provided in the Kaldi distribution. The acoustic models were trained using 109 h of speech: 79 h in Spanish (2006 TC-START http://tcstar.org/) and 30 in Galician (news database of Galicia, Transcrigal http://metashare.elda.org/repository/browse/transcrigal-db/72ee3974cbec11e181b50030482ab95203851f1f95e64c00b842977a318ef641/). The RTVE2018 database text files and text corpus of 90M words from several sources were used for language model training.
- G3-GTTS-EHU. Working group on software technologies, Universidad del País Vasco, Spain.
  This team participated with a commercial speech-to-text conversion system, with general purpose acoustic and language models. Only a primary system was submitted.
- G5-LIMECRAFT. Visiona Ingeniería de Proyectos, Madrid, Spain.
  This team participated with a commercial speech-to-text conversion system, with general purpose acoustic and language models. Only a primary system was submitted.

- G6-VICOMTECH-PRHLT [20]. VICOMTECH, Visual Interacion & Communication Technologies, Donostia, Spain and Pattern Recognition and PRHLT, Human Language Technologies Research Center, Universidad Politécnica de Valencia, Spain.
  G6-VICOMTECH-PRHLT submitted three systems. The primary system was an evolution of an already existing E2E (end-to-end) model based on DeepSpeech2 https://arxiv.org/abs/1512.02595, which was built using the three-fold augmented SAVAS https://cordis.europa.eu/project/rcn/103572/factsheet/en, Albayzin http://catalogue.elra.info/en-us/repository/browse/ELRA-S0089/, and Multext http://catalogue.elra.info/en-us/repository/browse/ELRA-S0060/ corpora for 28 epochs. For this challenge, it evolved with two new epochs using the same corpora in addition to the three-fold augmented nearly-perfectly aligned corpus obtained from the RTVE2018 dataset. A total of 897 h was used for training. The language model was a five-gram, trained with the text data from the open-set dataset. The first contrastive system was also an evolution of the already existing E2E model, but in this case, it evolved with one epoch using the three-fold augmented corpora used in the primary system and a new YouTube RTVE corpus. The duration of the total amount of training audio was 1488 h. The language model was a five-gram trained with the text data from the open-set dataset. The second contrastive system was composed of a bidirectional LSTM-HMM (long short term memory-hidden Markov model) acoustic model combined with a three-gram language model for decoding and a nine-gram language for re-scoring lattices. The acoustic model was trained with the same data as the primary system of the open-set condition. The language models were estimated with the RTVE2018 database text files and general news data from newspapers.
- G7-MLLP-RWTH [21]. MLLP, Machine Learning and Language Processing, Universidad Politécnica de Valencia, Spain and Human Language Technology and Pattern Recognition, RWTH Aachen University, Germany.
  G7-MLLP-RWTH submitted only a primary system. The recognition engine was an evolution of RETURNN https://github.com/rwth-i6/returnn and RASR https://www-i6.informatik.rwth-aachen.de/rwth-asr/, and it was based on a hybrid LSTM-HMM acoustic model. Acoustic modeling was done using a bidirectional LSTM network with four layers and 512 LSTM units in each layer. Three-thousand eight-hundred hours of speech transcribed from various sources (subtitled videos of Spanish and Latin American websites) were used for training the acoustic models. The language model for the single-pass HMM decoding was a five-gram count model trained with Kneser–Ney smoothing on a large body of text data collected from multiple publicly available sources. A lexicon of 325K words with one or more variants of pronunciation was used. Neither speaker, nor domain adaptation, nor model tuning were used.
- G14-SIGMA [22]. Sigma AI, Madrid, Spain.
  G14-SIGMA submitted only a primary system. The ASR system was based on the open-source Kaldi Toolkit. The ASR architecture consisted of the classical sequence of three main modules: an acoustic model, a dictionary or pronunciation lexicon, and an N-gram language model. These modules were combined for training and decoding using weighted finite state transducers (WFST). The acoustic modeling was based on deep neural networks and hidden Markov models (DNNHMM). To improve robustness mainly on speaker variability, speaker adaptive training (SAT) based on i-vectors was also implemented. Acoustic models were trained using 600 h from RTVE2018 (350 h of manual transcription), VESLIM https://ieeexplore.ieee.org/document/1255449/ (103 h), and OWNMEDIA (162 h of television programs). RTVE2018 database texts, news between 2015 and 2018, interviews, and subtitles were used to train the language model.
- G21-EMPHATIC [23]. SPIN-Speech Interactive Research Group, Universidad del País Vasco, Spain and Intelligent Voice, U.K.
  The G21-EMPHATIC ASR system was based on the open-source Kaldi Toolkit. The Kaldi Aspire recipe https://github.com/kaldi-asr/kaldi/tree/master/egs/aspire was used for building the DNNHMM acoustic model. Albayzin, Dihana http://universal.elra.info/product_info.php?

products_id=1421, CORLEC-EHU http://gtts.ehu.es/gtts/NT/fulltext/RodriguezEtal03a.pdf, and TC-START databases with a total of 352 h were used to train the acoustic models. The provided training and development audio files were subsampled to 8 kHz before being used in the training and testing processes. A three-gram LM base model trained with the transcripts of Albayzin, Dihana. CORLEC-EHU, TC-START, and a newspaper corpus (El País) was adapted using the selected training transcriptions of the RTVE2018 data.

4.1.2. Closed-Set Condition Systems

- G6-VICOMTECH-PRHLT [20]. VICOMTECH, Visual Interacion & Communication Technologies, Donostia, Spain and PRHLT, Pattern Recognition and Human Language Technologies Research Center, Universidad Politécnica de Valencia, Spain.
  G7-VICOMTECH-PRHLT submitted three systems. The primary system was a bidirectional LSTM-HMM based system combined with a three-gram language model for decoding and a nine-gram language model for re-scoring lattices. The training and development sets were aligned and filtered to get nearly 136 h of audio with transcription. The acoustic model was trained with a nearly perfectly aligned partition, which was three-fold augmented through the speed based augmentation technique obtaining a total of 396 h and 33 min of audio. The language models were estimated with the in-domain texts compiled from the RTVE2018 dataset. The first contrastive system was set up using the same configuration of the primary system, but the acoustic model was estimated using the three-fold augmented acoustic data of the perfectly aligned partition given a total of 258 h and 27 min of audio. The same data were used to build the second contrastive system, but it was an E2E recognition system that followed the architecture used for the open-set condition. The language model was a five-gram.
- G7-MLLP-RWTH [21]. MLLP, Machine Learning and Language Processing, Universidad Politécnica de Valencia, Spain and Human Language Technology and Pattern Recognition, RWTH Aachen University, Germany.
  G7-MLLP-RWTH submitted two systems. The recognition engine was the translectures-UPV toolkit (TLK) decoder [24]. The ASR consisted of a bidirectional LSTM-HMM acoustic model and a combination of both RNNLM and TV-show adapted n-gram language models. The training and development sets were aligned and filtered to get nearly 218 h of audio with transcription. For the primary system, all aligned data from train, dev1, and dev2 partitions, 218 h, were used for acoustic model training. For the contrastive system, only a reliable set of 205 h of the train and dev1 partitions were used for training the acoustic models. The language models were estimated with the in-domain texts compiled from the RTVE2018 dataset with a lexicon of 132 K words.
- G14-SIGMA [22]. Sigma AI, Madrid, Spain.
  G14-SIGMA submitted only a primary system. The system had the same architecture as the one submitted for the open-set conditions, but only 350 h of manual transcription of the training set were used for training the acoustic models. The language model was trained using the subtitles provided in the RTVE2018 dataset and manual transcriptions.

4.2. Speaker Diarization

A total of 30 different systems from nine participating teams were submitted. Six of them presented results in the closed-set condition and five in the open-set condition. The most relevant characteristics of each system are presented in terms of the diarization technology and the data used for training models.

4.2.1. Open-Set Condition Systems

- G1-GTM-UVIGO. Multimedia Technologies Group, Universidad de Vigo, Spain.
  A pre-trained deep neural network http://kaldi-asr.org/models.html was used with Kaldi and data from VoxCeleb1 and VoxCeleb2 databases http://www.robots.ox.ac.uk/~vgg/data/

voxceleb/. This DNN mapped variable length speech segments into fixed dimension vectors called x-vectors. The strategy followed for diarization consisted of three main stages. The first stage was the extraction of the x-vector and grouping using the clustering algorithm "Chinese Whispers". In the second stage, each of the audio segments was processed to extract one or more x-vectors using a sliding window of 10 s with half a second of displacement between successive windows. These vectors were grouped using the clustering algorithm "Chinese Whispers" obtaining the clusters that defined the result of the diarization. Finally, a music/non-music discriminator based on i-vectors and a logistic regression model were applied to eliminate those audio segments that were highly likely to correspond to music. This discriminator was also trained with external data.

- G11-ODESSA [25]. EURECOM, LIMSI, CNRS, France.
  A primary system resulting from the combination at a similarity matrix level of three systems, one trained according to the closed-set condition and two trained with two external databases (NIST SRE (Speaker Recognition Evaluation) and Voxceleb), was submitted. The first contrastive system used one second uniform segmentation, x-vector representation trained on NIST SRE data, and agglomerative hierarchical clustering (AHC). The second contrastive system was the same as the one for the closed-set, but where the training data were replaced with the Voxceleb data.
- G20-STAR-LAB [26]. STAR Lab, SRI International, USA.
  The training signals were extracted from the databases NIST SRE https://www.nist.gov/itl/iad/mig/speaker-recognition 2004–2008, NIST SRE 2012, Mixer6 https://catalog.ldc.upenn.edu/LDC2013S03, Voxceleb1, and Voxceleb2. Augmentation of data was applied using four categories of degradations including music, noise at a 5 dB signal-to-noise ratio, compression, and low levels of reverb. STAR-Lab used the embeddings and diarization system developed for the speaker recognition competition NIST 2018 [27]. It incorporated modifications in the detection of voice activity and in the calculation of embeddings for speaker recognition. The differences between systems, primary and contrast, were found in the different parameters used in the voice activity detection system.
- **G21-EMPHATIC** [28]. SPIN-Speech Interactive Research Group, Universidad del País Vasco, Spain, Intelligent Voice, U.K.
  The Switchboard corpora databases (LDC2001S13, LDC2002S06, LDC2004S07, LDC98S75, LDC99S79) and NIST SRE 2004–2010 were used for training. Data augmentation was used to provide a greater diversity of acoustic environments. The representation of speaker embeddings was obtained through an end-to-end model using convolutional (CNN) and recurrent (LSTM) networks.
- G22-JHU [29]. Center for Language and Speech Processing, Johns Hopkins University, USA.
  Results were presented with different databases for training. Voxceleb1 and Voxceleb2 were used with and without data augmentation, SRE12-micphn, MX6-micph, and SITW-dev-core, Fisher database, Albayzin2016, and RTVEDB2018. In relation to the diarization system, several systems based on four different types of embeddings extractors: x-vector-basic, x-vector-factored, i-vector-basic, and bottleneck features (BNF) i-vector were used. All the systems followed the structure: parameter extraction, voice activity detector, embeddings extraction, probabilistic linear discriminant analysis (PLDA) scoring, fusion and grouping of speakers.

4.2.2. Closed-Set Condition Systems

- G4-VG [30]. Voice Group, Advanced Technologies Application Center-CENATAV, Cuba.
  The submitted systems used a classic structure based on Bayesian information criterion (BIC) segmentation, hierarchical agglomerative grouping, and re-segmentation by hidden Markov models. The toolbox S4D https://projets-lium.univ-lemans.fr/s4d/ was used. The difference between the submitted systems was a feature extraction method ranging from classic ones as Mel frequency cepstral coefficients (MFCC), linear frequency cepstral coefficients (LFCC), and linear

prediction cepstral coefficients (LPCC), to new ones such as mean Hilbert envelope coefficients [31], medium duration modulation coefficients, and power normalization cepstral coefficients [32].

- G8-AUDIAS-UAM [33]. Audio, Data Intelligence and Speech, Universidad Autónoma de Madrid, Spain.
  Three different systems were submitted, two based on DNN based embeddings using an architecture based on bidirectional LSTM recurrent neural network (primary and first contrastive systems) and a third one based on the classical model of total variability (second contrastive system).
- G10-VIVOLAB [34]. ViVoLAB, Universidad de Zaragoza, Spain.
  The system was based on the use of i-vectors with PLDA. The i-vectors were extracted from the audio in accordance with the assumption that each segment represented the intervention of a speaker. The hypothesis of diarization was obtained by grouping the i-vectors with a fully Bayesian PLDA. The number of speakers was decided by comparing multiple hypotheses according to the information provided by the PLDA. The primary system performed unsupervised PLDA adaptation, while the contrastive one did not.
- G11-ODESSA [25]. EURECOM, LIMSI, CNRS, France.
  The primary system was the fusion at a diarization hypothesis level of the two contrastive systems. The first contrastive system used ICMCfeatures (infinite impulse response—constant Q, Mel-frequency cepstral coefficients), one second uniform segmentation, binary key (BK) representation, and AHC, while the second one used MFCC features, bidirectional LSTM based speaker change detection, triplet-loss neural embedding representation, and affinity propagation clustering.
- G19-EML [35]. European Media Laboratory GmbH, Germany.
  The submitted diarization system was designed primarily for on-line applications. Every 2 s, it made a decision about the identity of the speaker without using future information. It used speaker vectors based on the transformation of a Gaussian mixture model (GMM) supervector.
- G22-JHU [29]. Center for Language and Speech Processing, Johns Hopkins University, USA.
  The system submitted in the closed-set condition was similar to that of the open-set condition with the difference that in this case, only embeddings based on i-vectors were used.

4.3. Multimodal Diarization

A total of 10 different systems from four participating teams were submitted.

System Descriptions

- G1-GTM-UVIGO [36]. Multimedia Technologies Group, Universidad de Vigo, Spain.
  The proposed system used state-of-the-art algorithms based on deep neural networks for face detection and identification and speaker diarization and identification. Monomodal systems were used for faces and speaker, and the results of each system were fused to better adjust the speech of speakers.
- G2-GPS-UPC [37]. Signal processing group, Universidad Politécnica de Cataluña, Spain.
  The submitted system consisted of two monomodal systems and a fusion block that combined the outputs of the monomodal systems to refine the final result. The audio and video signal were processed independently, and they were merged assuming there was a temporal correlation between the speaker's speech and face, that there were talking characters, his/her face did not appear, and faces that appeared, but did not speak.
- G9-PLUMCOT [38]. LIMSI, CNRS, France.
  The submitted systems (primary and two contrastive) made use of technologies based on monomodal neural networks: segmentation of the speaker, embeddings of speakers, embeddings

of faces, and detection of talking faces. The PLUMCOT system tried to optimize various hyperparameters of the algorithms by jointly using visual and audio information.

- G11-ODESSA [38]. EURECOM, LIMSI, CNRS, France, IDIAP, Switzerland.
  The system submitted by ODESSA was the same as the one presented by PLUMCOT, with the difference that in this case, the diarization systems were totally independent, and each one was optimized in a monomodal way.

5. Results

This section is dedicated to presenting the results obtained in the three challenges by the participating teams. A brief description of the teams and their systems is found in Section 4.

5.1. Speech-to-Text Evaluation

A total of 18 systems was submitted by seven teams, 12 systems to the open-set condition and six to the closed-set one. Results are presented for the open and closed conditions.

5.1.1. Open-Set Condition Results

Table 5 presents the overall results in the open-set condition by show and system. Results are given in terms of the average word error rates calculated over all the episodes of a show and the average over all the shows for a system. The best system, presented by the G7-MLLP-RWTH team, showed a WER of 16.45% using a hybrid LSTM-HMM ASR system. The second place corresponded to the system presented by the G14-SIGMA team, with a WER of 18.63% using the Kaldi toolkit. The first fully commercial and general purpose system was in third position, G5-LIMECRAFT team, with 20.92% of WER. The G6-VICOMTECH-PRHLT team achieved 24.52% of WER with the DeepSpeech2 system, an end-to-end system based entirely on deep neural networks. The second commercial system, G3-EHU, achieved 28.72% of WER, and finally, the other two teams that used Kaldi, G1-GTM-VIGO and G21-EMPHATIC, obtained 29.27% and 31.61% of WER, respectively.

If we compare the audio and text resources used to train the systems, except for the commercial systems whose information was not provided, the G7-MLLP-RHTW team was the one that used the most resources for training acoustic and language models, followed by the G14-SIGMA and G6-VICOMTECH-PRHLT teams. Furthermore, it should be taken into account that the G14-SIGMA acoustic models were trained using 350 h of the training partition with manual transcription, whereas G6-VICOMTECH-PRHLT used aligned and filtered data from the training and development sets. The correlation between performance and training resources was clear: the more resources, the better the results.

Regarding the WER per TV shows, the variance across shows was quite high. The WER per show varied from 7.43% to 34.45% for the most accurate system. The LA24H (Latinoamérica en 24H) and EC (España en Comunidad) shows were the ones with lower WER. A priori, the good results obtained by LA24H were not expected, as the data contained Spanish accents from Latin America. However, most of the time, the LA24H show contained a high quality voice-over in terms of acoustic environment and Spanish pronunciation, which can somehow explain the good result. The worst results were given by the "Dicho y Hecho" quiz show due to the acoustic environment and speech inflections of the participants. Furthermore, the WER variability among episodes of the same show was high. Figure 1 shows a boxplot of the WER variability per show for the best system.

Table 6 shows the overall text normalized WER (TNWER). The reference and test text were normalized in terms of removing stop-words and changing each word by its lemma. The relative WER reduction was about 5% when using text normalization to compute WER.

Only two teams submitted results with punctuation marks. Table 7 shows the overall results in terms of PWER (see Section 3.1.2) for the submitted systems. There were four systems submitted with periods and three with periods and commas. In all the cases, the performance was degraded mainly by the increase in the number of deletions and insertions related to periods and commas.

Team G6-VICOMTECH-PRHLT obtained much better results with the second contrastive system than the primary. The primary system used an E2E approach, but the second contrastive system was a bidirectional LSTM-HMM system with an n-gram language model. The training data were the same for primary and contrastive systems.

Table 5. Speech-to-text open-set condition. Word error rate (WER) (%) per TV show and team. Team and system descriptions are in Section 4.1.1. See Table 4 for show descriptions. P, primary; C#, contrastive #). Best results by TV show are in bold.

Team	**G1**		**G3**	**G5**	**G6**			**G7**	**G14**	**G21**		
System	P	C1	P	P	P	C1	C2	P	P	P	C1	C2
TV Show												
AFI	29.79	30.39	19.72	16.35	22.37	28.47	25.99	**15.91**	17.65	28.22	31.48	29.57
AV	54.67	54.68	47.13	39.97	40.49	48.36	42.17	**23.94**	28.90	39.14	54.75	50.21
DH	56.53	56.58	59.18	41.50	49.44	56.77	51.30	**34.45**	43.06	51.24	58.50	53.82
EC	21.86	22.54	17.99	15.59	17.64	23.68	20.81	**11.38**	13.54	22.19	25.60	23.32
LA24H	14.75	15.94	15.41	8.23	11.87	16.69	12.74	**7.43**	9.43	14.70	16.53	14.99
LM	36.74	37.58	38.35	27.10	31.72	44.69	34.40	**21.94**	23.96	45.94	47.70	46.43
LT24HTer	27.37	28.57	28.37	20.61	23.34	31.14	24.82	18.97	**17.41**	32.90	39.18	37.29
SG	25.43	27.28	31.47	19.66	22.81	33.82	22.65	15.97	**14.77**	21.32	21.10	20.16
Overall WER	29.27	30.19	28.72	20.92	24.52	33.00	26.66	**16.45**	18.63	31.61	35.80	33.90

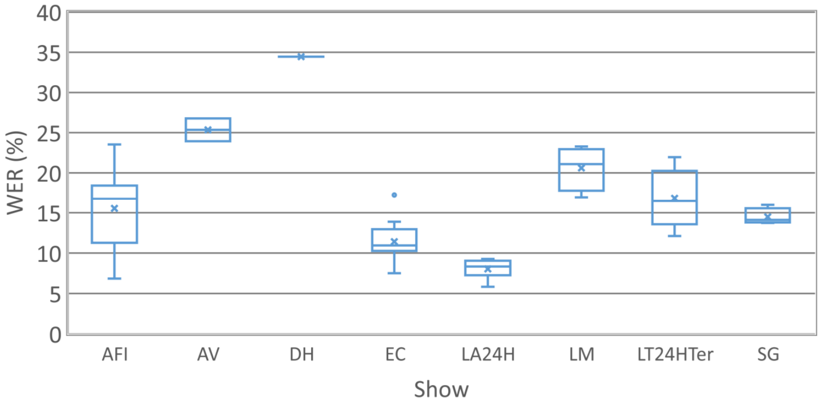

Figure 1. Boxplot with means (x) of the WER by shows for the best system in the open-set condition. See Table 4 for show descriptions.

Table 6. Speech-to-text open-set condition. Text normalized word error rate (TNWER) (%) for the primary systems per show and team. Team and system descriptions are in Section 4.1.1. See Table 4 for show descriptions. Best results by TV show are in bold.

Team	**G1**	**G3**	**G5**	**G6**	**G7**	**G14**	**G21**
TV Show							
AFI	27.25	18.35	15.35	21.41	**14.06**	16.04	26.11
AV	53.38	46.67	39.69	39.95	**25.93**	28.69	41.83
DH	56.82	60.05	42.76	49.7	**34.62**	43.32	51.16
EC	19.72	16.75	14.36	16.39	**10.47**	11.95	20.89
LA24H	13.11	14.56	7.49	11.14	**7.27**	8.21	13.97
LM	34.65	38.00	26.26	30.93	**19.79**	22.59	41.53
LT24HTer	25.23	27.76	19.97	22.31	15.75	**15.42**	29.65
SG	25.01	31.86	20.1	23.45	**15.11**	15.20	20.95
Overall TNWER	27.36	28.1	20.21	23.69	**15.69**	17.26	29.59
Overall WER	29.27	28.72	20.92	24.52	**16.45**	18.63	31.61

Table 7. Speech to text open-set condition. WER (%) and punctuation marks evaluation (PWER) (%) for the primary systems per team. Team and system descriptions are in Section 4.1.1.

Punc.Marks	**Periods**		**Periods**			**Periods and Commas**	
Team	G5		G6			G6	
System	P	P	C1	C2	P	C1	C2
Overall WER	20.92	24.52	33.00	26.66	24.52	33.00	26.66
Overall PWER	26.04	29.75	29.19	28.12	33.00	33.59	31.59

5.1.2. Closed-Set Condition Results

Regarding the closed-set condition results, Table 8 presents the overall results of the three primary systems. The best results, with a WER of 19.57%, were obtained by the G14-SIGMA system. The G14-SIGMA closed-set condition system was similar to the one used in the open-set condition; the difference was in the data used for training the acoustic and language models. In the open-set condition, acoustic and language models were trained with the data used in the closed condition augmented with an additional 265 h of audio and text coming from news, interviews, and subtitles. This data augmentation allowed a reduction of the WER from 19.57% to 17.26%. It is interesting to compare results among the three systems in terms of the data used for training. The main difference was the way they obtained correctly transcribed audio from the training and development partitions. G14-SIGMA used manual transcription of 350 h of the train partition for both acoustic and language models. However, G6-VICOMTECH-PRHLT and G7-MLLP-RHTW used automatic aligned and filtered audio from the train and development partitions for training acoustic models, a total of 136 h and 218 h, respectively, and all text files given in the RTVE2018 database for training language models. Although the three systems were using different ASR toolkits, the results showed a high correlation between the amount of training data obtained from the training set and the performance: the more training data, the better the results. It is valuable to highlight that the filtering techniques used by G6-VICOMTECH-PRHLT and G7-MLLP-RHTW degraded the performance by less than 3% compared to the human-transcription used by G14-SIGMA.

In terms of the TNWER (Table 9), it is interesting to note that G14-SIGMA almost got the same TNWER (17.90%) as in the open-set condition (17.26%), which means that most of the errors in the closed-set conditions were related to stop-words, verbal conjugations, and word number or gender. A comparison between open-set and closed-set conditions for the G6-VICOMTECH-PRHLT and G7-MLLP-RHTW systems was not possible as they used different ASR architectures.

Table 8. Speech-to-text closed-set condition. WER(%) per TV show and team. Team and system descriptions are in Section 4.1.2. See Table 4 for show descriptions. Best results by TV show are in bold.

Team		**G6**			**G7**	**G14**
System	P	C1	C2	P	C1	P
TV Show						
AFI	24.22	25.01	25.99	25.39	25.29	**19.75**
AV	33.94	37.75	42.17	36.17	37.19	**27.75**
DH	45.62	48.36	51.30	50.82	50.35	**43.50**
EC	16.70	17.27	20.81	16.68	16.56	**15.63**
LA24H	**10.47**	10.73	12.74	12.04	11.96	11.25
LM	28.28	29.58	34.40	26.68	26.60	**23.96**
LT24HTer	20.80	21.38	24.82	19.15	19.29	**18.01**
SG	17.7	18.92	22.65	18.79	18.75	**15.43**
Overall WER	22.22	23.16	26.66	21.98	21.96	**19.57**

Table 9. Speech to text closed-set condition. WER (%) and TNWER (%) for the primary systems per team. Team and system descriptions are in Section 4.1.2. Best results by TV show are in bold.

Team	G6	G7	G14
Overall WER	22.22	21.98	**19.57**
Overall TNWER	20.71	19.75	**17.90**

5.2. Speaker Diarization Evaluation

Nine teams participated in the speaker diarization task submitting 26 systems, 13 for each condition, open-set and closed-set.

5.2.1. Open-Set Condition Results

Five teams participated in the open-set condition evaluation. Four teams submitted results on time, but G1-GTM-UVIGO made a late submission. This late submission was not taken into account for the challenge ranking, but we included them in this review, as their results were very impressive. Table 10 presents results for each submitted system by team. G1-GTM-UVIGO obtained a DER of 11.4%, which is almost half of the next system, the contrastive C1 system from the G11-ODESSA team with a DER of 20.3%. The most significant difference between the G1-GTM-UVIGO systems and the rest of the systems was the low speaker error, as G1-GTM-UVIGO obtained 6.6% and the next obtained 16.8%. Figure 2 presents the average per show of the estimated number of speakers of the best systems of each team. Considering that the correct average per show of the number of speakers in the database was 27, G1-GTM-VIGO made a close estimate of 28 speakers per show on average. The rest of the systems underestimated substantially the number of speakers present in each show. This estimation of the number of speakers for each show was a key point that allowed us to explain the good results of the G1-GTM-UVIGO in this evaluation.

On the other hand, Table 11 presents the DER obtained by each participant for each different show. The lower DER was obtained in "La Tarde en 24H Tertulia", which is a talk show about politics and the economy with 20 speakers on average, good quality audio, and no overlapping speech. The rest of the shows include set and outdoor sections with a higher number of speakers: 65 for LM (La Mañana), 34 for LA24H (Latinoamérica en 24H) and 29 for EC (España en Comunidad).

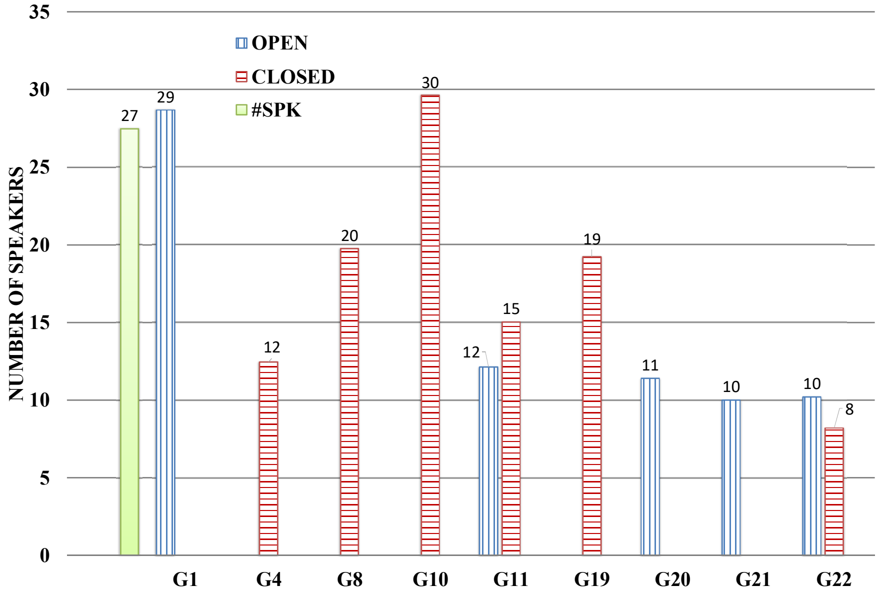

Figure 2. Average per show of the estimated number of speakers of the primary system of each team. #SPK is the real average number of speakers. Team and system descriptions in Section 4.2.1.

Table 10. Open-set condition speaker diarization: diarization error rate (DER) (%), missed speech (%), false speech (%), and speaker error (%) per team. Team and system descriptions are in Section 4.2.1. Best results are in bold.

Team	**G1**			**G11**			**G20**			**G21**	**G22**		
System	P	C1	C2	P	C1	C2	P	C1	C2	P	P	C1	C2
DER	**11.4**	11.7	12.7	25.9	20.3	36.7	30.8	31.8	33.3	30.96	28.6	28.2	31.4
Missed Speech	1.1	1.9	1.2	0.7	0.7	0.7	0.7	0.7	**0.6**	0.9	2.4	2.4	2.4
False Speech	3.7	3.2	3.7	2.8	2.8	2.8	3.1	3.7	4.5	4.8	**1.3**	1.3	1.3
Speaker Error	**6.6**	6.6	7.8	22.4	16.8	33.2	26.9	27.4	28.2	25.2	24.9	24.5	27.7

Table 11. Open-set condition speaker diarization: DER (%) for the best systems per team and TV show. Team and system descriptions are in Section 4.2.1.

Team	**G1**	**G11**	**G20**	**G21**	**G22**
System	P	C1	P	P	C1
TV Show					
EC	13.1	27.4	37.7	40.9	38.6
LA24H	15.0	29.3	39.5	36.7	34.3
LM	16.9	24.1	48.7	45.3	35.9
LT24HTer	7.8	9.9	18.4	18.2	15.6
DER	11.4	20.3	30.7	30.9	28.2

5.2.2. Closed-set Condition Results

Six teams participated in the closed-set condition. Table 12 shows the results of all the submitted systems. The best result was obtained by the team G10-VIVOLAB with a DER of 17.3% for the primary system. The second best result was obtained by the team G4-VG with a DER of 25.4% with the second contrastive system. One of the most noticeable differences between the more accurate system and the rest was the speaker error term, which was substantially lower than the rest. This lower speaker error term was correlated with the more accurate estimation of the number of speakers, as can be seen in Figure 2, where the average of the estimated number of speakers per show is presented. It is interesting to compare the open-set and closed-set conditions presented by G11-ODESSA. Contrastive System 2 (C2) was the same for both conditions, the only difference was the training dataset, Voxceleb data for the open-set condition and the RTVE2018 dataset for the closed-set condition. The results were very similar with a slight improvement on the open-set condition. Regarding the results per show, Table 13 presents the diarization errors for the best system of each team, showing a similar trend as in the open-set condition.

Table 12. Closed-set condition speaker diarization: DER (%), missed speech (%), false speech (%), and speaker error (%) per team. Team and system descriptions are in Section 4.2.2. Best results are in bold.

Team	**G4**			**G8**			**G10**		**G11**			**G19**	**G22**
System	P	C1	C2	P	C1	C2	P	C1	P	C1	C2	P	P
DER	26.7	26.5	25.4	34.6	31.4	28.7	**17.3**	17.8	26.6	30.2	37.6	26.6	39.1
Missed Speaker	**0.4**	0.4	0.4	3.1	2.5	4.1	1.1	1.1	0.7	0.7	0.7	1.1	2.4
False Speaker	4.8	4.8	4.8	3.1	3.2	3.5	2.5	2.5	2.8	2.9	2.8	3	**1.3**
Speaker Error	21.5	21.3	20.2	28.4	25.7	21.1	**13.7**	14.2	23.1	26.6	34.1	22.5	35.4

Table 13. Closed-set condition speaker diarization: DER (%), missed speech (%), false speech (%), and speaker error (%) for the best systems per team. Team and system descriptions are in Section 4.2.2. Best results by TV show are in bold.

Team	**G4**	**G8**	**G10**	**G11**	**G19**	**G22**
System	C2	C2	P	P	P	P
TV Show						
EC	34.8	27.8	**17.1**	37.6	30.3	47.0
LA24H	30.7	31.3	**18.2**	31.1	30.7	44.6
LM	**29.1**	36.0	41.2	33.7	35.3	52.4
LT24HTer	14.9	27.6	**13.3**	14.6	20.2	27.9
Overall DER	25.4	28.7	**17.3**	26.6	26.6	39.1
Missed Speech	**0.4**	4.1	1.1	0.7	1.1	2.4
False Speech	4.8	3.5	**2.5**	2.8	3	1.3
Speaker Error	20.2	21.1	**13.7**	23.1	22.5	35.4

5.3. Multimodal Diarization Evaluation

Table 14 shows the results of the multimodal diarization evaluation in terms of the overall diarization error and the diarization error for each modality (speaker and face). These results are presented for each submitted system and for each show contained in the test set. Note that the number of different characters to identify was 39. Three out of the four systems that were submitted for this task obtained very similar results in terms of the overall DER, ranging between 23% and 30%. The best result was obtained by the system presented by G9-PLUMCOT with 23.4% of overall DER. It is worth remarking that thanks to the optimization of speaker diarization using face information, the team was able to reduce the DER from 29.1%, obtained by independent optimization, to 23.4%. Note that G11-ODESSA and G9-PLUMCOT used the same face system for this evaluation. The results obtained by G2-GPS-UPC were significantly worse due to the poor adjustment of the speaker diarization system. Except for G2-GPS-UPC, the speech modality achieved significantly better results than the face modality. The G1-GTM-UVIGO speaker diarization system was the one that obtained the best results, 17.3%, followed by the G11-PLUMCOT system with 17.6%. For the face modality, the system used by G11-ODESSA and G9-PLUMCOT obtained the best results followed by the G2-GPS-UPC system. Regarding the results obtained by the participants for the different shows included in the evaluation, "La Mañana" (LM-20170103) presented the highest difficulty for both face and speech modalities mainly due to the fact that it is a live show with a large amount of overlapping speech and outdoor scenes. On the other side, "La tarde en 24 Horas, Tertulia" (LT24HTer-20180222 and LT24HTer-20180223) presented TV studio set scenes mostly, less overlapping speech, and only five or less speakers and characters per show.

Table 14. Multimodal diarization: DER (%), missed speech/face (%), false speaker/face (%) and speaker/face error (%) for the best systems. Team and system descriptions in Section 4.3. Best results are in bold.

Team	**G1**		**G2**		**G9**		**G11**	
Modality/System	Face/P	SPK/P	Face/P	SPK/C1	Face/P	SPK/P	Face/P	SPK/C1
Missed Speech/Face	28.6	3.0	22.6	**1.1**	**12.1**	1.2	**12.1**	**1.1**
False Speech/Face	**5.2**	9.3	**4.7**	14.1	12.2	12.4	12.2	12.6
Error Speaker/Face	**0.9**	5.0	2.3	47.4	**4.9**	4.0	**4.9**	15.2
Overall DER	34.7	**17.3**	29.6	62.6	**29.2**	17.6	**29.2**	28.9
Multimodal DER	26.0		46.1		**23.4**		29.1	
TV Show								
LM-20170103	57.7	**35.5**	43.0	74.6	**44.3**	43.7	**44.3**	63.1
LT24HTer-20180222	19.8	9.0	19.4	46.4	**17.3**	**3.2**	**17.3**	22.8
LT24HTer-20180223	**19.0**	8.2	21.5	63.1	20.2	**6.3**	20.2	6.6

6. Conclusions

The IberSpeech-RTVE 2018 challenge was a new edition of the Albayzin evaluation series focused on speech-to-text transcription, speaker diarization, and multimodal diarization of TV broadcast content. A broad international participation was achieved with 18 teams, five of them participating in more than one task. A new dataset named the RTVE2018 database was released containing more than 500 h of TV shows.

Seven teams participated in the speech-to-text task submitting 18 different systems, 12 for the open-set condition and 6 for the closed-set condition. The evaluation was carried out over eight different shows covering a broad range of acoustic conditions. The performance of the submitted systems presented a high variability across the different shows that were included in the test set, ranging between 7.43% and 34.45% for the best system in the open-set condition. The most accurate system obtained an overall WER of 16.56%. The closed-set condition obtained worse results due to the difficulty of getting correctly transcribed audio out of the training dataset. The best result, 19.57% WER, was obtained training the acoustic model with 350 h of human-transcribed audio. Text normalization removing stop-words and lemmatization gave a small improvement in terms of WER.

In the speaker diarization task, nine teams submitted 26 different systems, 13 for each condition, closed-set and open-set. The evaluation was carried out over four different shows covering a broad area of acoustic conditions, number of different speakers, and amount of overlapping speech. The best systems in the open-set and closed-set conditions obtained 11.4% and 17.3% of DER, respectively. There was a big gap in terms of DER between the best system and the rest in both conditions. The success was mainly due to the accuracy in the estimation of the number of speakers for each show. The rest of the submitted system made a clear underestimation of the number of speakers.

The multimodal diarization challenge achieved the participation of four teams. The test set was composed of two different shows, one episode of a live show (La Mañana) and two episodes of a talk show about politics and the economy. The task consisted of identifying voices and faces that belonged to persons from the closed-set of 39 different characters. The most accurate system obtained a multimodal DER of 23.4%. In terms of individual modalities, face and speech, the best results were 17.6% of speaker DER and 29.2% of face DER. In the evaluation, always the speech modality obtained lower DER values than the face modality.

We plan to continue organizing a new Albayzin evaluation edition in the next IberSpeech conference in 2020. An extension of the database with new annotated audiovisual material for the development and evaluation of new audiovisual technologies is under preparation.

Author Contributions: Conceptualization, E.L., A.O., A.M., V.B.-G., C.P., M.G., and A.d.P.; formal analysis, E.L., A.O., and A.M.; resources, V.B.-G., C.P., M.G., and A.d.P.; writing, original draft preparation, E.L., A.O., and A.M.; writing, review and editing, E.L., A.O., A.M., and V.B.-G.

Funding: This work was supported by the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the project TIN2017-85854-C4-1-R, the Government of Aragón (Reference Group T36_17R) and co-financed with Feder 2014–2020 "Building Europe from Aragón", Radio Televisión Española through the RTVE Chair at the University of Zaragoza, and Red Temática en Tecnologías del Habla 2017, (TEC2017-90829-REDT) founded by Ministerio de Ciencia, Innovación y Universidades.

Acknowledgments: We gratefully acknowledge the support of the IberSpeech 2018 organizers.

Conflicts of Interest: The authors declare no conflict of interest.

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, Chantilly, VA, USA, 2–5 February 1997.
2. Graff, D. An overview of Broadcast News corpora. *Speech Commun.* **2002**, *37*, 15–26. [CrossRef]

3. Galliano, S.; Geoffrois, E.; Gravier, G.; Bonastre, J.; Mostefa, D.; Choukri, K. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), Genoa, Italy, 22–28 May 2006; pp. 315–320.
4. Galliano, S.; Gravier, G.; Chaubard, L. The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In Proceedings of the Interspeech, Brighton, UK, 6–10 September 2009; pp. 2583–2586.
5. Mostefa, D.; Hamon, O.; Choukri, K. Evaluation of Automatic Speech Recognition and Spoken Language Translation within TC-STAR: Results from the first evaluation campaign. In Proceedings of the LREC'06, Genoa, Italy, 22–28 May 2006.
6. Butko, T.; Nadeu, C. Audio segmentation of broadcast news in the Albayzin-2010 evaluation: Overview, results, and discussion. *EURASIP J. Audio Speech Music Process.* **2011**. [CrossRef]
7. Zelenák, M.; Schulz, H.; Hernando, J. Speaker diarization of broadcast news in Albayzin 2010 evaluation campaign. *EURASIP J. Audio Speech Music Process.* **2012**. [CrossRef]
8. Ortega, A.; Castan, D.; Miguel, A.; Lleida, E. The Albayzin-2012 audio segmentation evaluation. In Proceedings of the IberSpeech, Madrid, Spain, 21–23 November 2012.
9. Ortega, A.; Castan, D.; Miguel, A.; Lleida, E. The Albayzin-2014 audio segmentation evaluation. In Proceedings of the IberSpeech, Las Palmas de Gran Canaria, Spain, 19–21 November 2014.
10. Castán, D.; Ortega, A.; Miguel, A.; Lleida, E. Audio segmentation-by-classification approach based on factor analysis in broadcast news domain. *EURASIP J. Audio Speech Music Process.* **2014**, *34*. [CrossRef]
11. Ortega, A.; Viñals, I.; Miguel, A.; Lleida, E. The Albayzin-2016 Speaker Diarization Evaluation. In Proceedings of the IberSpeech, Lisbon, Portugal, 23–25 November 2016.
12. 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 2015, Scottsdale, AZ, USA, 13–17 December 2015; pp. 687–693. [CrossRef]
13. Ali, A.M.; Bell, P.; Glass, J.R.; Messaoui, Y.; Mubarak, H.; Renals, S.; Zhang, Y. The MGB-2 challenge: Arabic multi-dialect broadcast media recognition. In Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, USA, 13–16 December 2016; pp. 279–284. [CrossRef]
14. Ali, A.; Vogel, S.; Renals, S. Speech recognition challenge in the wild: Arabic MGB-3. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2017), Okinawa, Japan, 16–20 December 2017; pp. 316–322. [CrossRef]
15. Versteegh, M.; Thiollière, R.; Schatz, T.; Cao, X.; Anguera, X.; Jansen, A.; Dupoux, E. The zero resource speech challenge 2015. In Proceedings of the INTERSPEECH 2015—16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015; pp. 3169–3173.
16. Dunbar, E.; Cao, X.; Benjumea, J.; Karadayi, J.; Bernard, M.; Besacier, L.; Anguera, X.; Dupoux, E. The zero resource speech challenge 2017. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2017), Okinawa, Japan, 16–20 December 2017; pp. 323–330. [CrossRef]
17. Eskevich, M.; Aly, R.; Racca, D.N.; Ordelman, R.; Chen, S.; Jones, G.J.F. The Search and Hyperlinking Task at MediaEval 2014. In Proceedings of the Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Catalunya, Spain, 16–17 October 2014.
18. Zelenak, M.; Schulz, H.; Hernando, J. Albayzin 2010 Evaluation Campaign: Speaker Diarization. In Proceedings of the VI Jornadas en Tecnologías del Habla—FALA 2010, Vigo, Spain, 10–12 November 2010.
19. Docío-Fernández, L.; García-Mateo, C. The GTM-UVIGO System for Albayzin 2018 Speech-to-Text Evaluation. In Proceedings of the IberSPEECH, Barcelona, Spain, 21–23 November 2018; pp. 277–280. [CrossRef]
20. Arzelus, H.; Alvarez, A.; Bernath, C.; García, E.; Granell, E.; Martinez Hinarejos, C.D. The Vicomtech- PRHLT Speech Transcription Systems for the IberSPEECH-RTVE 2018 Speech to Text Transcription Challenge. In Proceedings of the IberSPEECH, Barcelona, Spain, 21–23 November 2018; pp. 267–271. [CrossRef]
21. Jorge, J.; Martínez-Villaronga, A.; Golik, P.; Giménez, A.; Silvestre-Cerdà, J.A.; Doetsch, P.; Císcar, V.A.; Ney, H.; Juan, A.; Sanchis, A. MLLP-UPV and RWTH Aachen Spanish ASR Systems for the IberSpeech-RTVE 2018 Speech-to-Text Transcription Challenge. In Proceedings of the IberSPEECH, Barcelona, Spain, 21–23 November 2018; pp. 257–261. [CrossRef]

22. Perero-Codosero, J.M.; Antón-Martín, J.; Tapias Merino, D.; López-Gonzalo, E.; Hernández-Gómez, L.A. Exploring Open-Source Deep Learning ASR for Speech-to-Text TV program transcription. In Proceedings of the IberSPEECH, Barcelona, Spain, 21–23 November 2018; pp. 262–266. [CrossRef]
23. Dugan, N.; Glackin, C.; Chollet, G.; Cannings, N. Intelligent Voice ASR system for IberSpeech 2018 Speech to Text Transcription Challenge. In Proceedings of the IberSPEECH, Barcelona, Spain, 21–23 November 2018; pp. 272–276. [CrossRef]
24. Del Agua, M.; Giménez, A.; Serrano, N.; Andrés-Ferrer, J.; Civera, J.; Sanchis, A.; Juan, A. The translectures-UPV toolkit. In *Advances in Speech and Language Technologies for Iberian Languages*; Springer: Cham, Switzerland, 2014; pp. 269–278.
25. Patino, J.; Delgado, H.; Yin, R.; Bredin, H.; Barras, C.; Evans, N. ODESSA at Albayzin Speaker Diarization Challenge 2018. In Proceedings of the IberSPEECH, Barcelona, Spain, 21–23 November 2018; pp. 211–215. [CrossRef]
26. Castan, D.; McLaren, M.; Nandwana, M.K. The SRI International STAR-LAB System Description for IberSPEECH-RTVE 2018 Speaker Diarization Challenge. In Proceedings of the IberSPEECH, Barcelona, Spain, 21–23 November 2018, pp. 208–210. [CrossRef]
27. McLaren, M.; Ferrer, L.; Castan, D.; Nandwana, M.; Travadi, R. The sri-con-usc nist 2018 sre system description. In Proceedings of NIST 2018 Speaker Recognition Evaluation, Athens, Greece, 16–17 December 2018.
28. Khosravani, A.; Glackin, C.; Dugan, N.; Chollet, G.; Cannings, N. The Intelligent Voice System for the IberSPEECH-RTVE 2018 Speaker Diarization Challenge. In Proceedings of the IberSPEECH, Barcelona, Spain, 21–23 November 2018; pp. 231–235. [CrossRef]
29. Huang, Z.; García-Perera, L.P.; Villalba, J.; Povey, D.; Dehak, N. JHU Diarization System Description. In Proceedings of the IberSPEECH, Barcelona, Spain, 21–23 November 2018; pp. 236–239. [CrossRef]
30. Campbell, E.L.; Hernandez, G.; Calvo de Lara, J.R. CENATAV Voice-Group Systems for Albayzin 2018 Speaker Diarization Evaluation Campaign. In Proceedings of the IberSPEECH, Barcelona, Spain, 21–23 November 2018; pp. 227–230. [CrossRef]
31. Sadjadi, S.O.; Hansen, J.H. Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification. *Speech Commun.* **2015**, *72*, 138–148. [CrossRef]
32. Kim, C.; Stern, R.M. Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 1315–1329. [CrossRef]
33. Lozano-Diez, A.; Labrador, B.; de Benito, D.; Ramirez, P.; Toledano, D.T. DNN-based Embeddings for Speaker Diarization in the AuDIaS-UAM System for the Albayzin 2018 IberSPEECH-RTVE Evaluation. In Proceedings of the IberSPEECH, Barcelona, Spain, 21–23 November 2018; pp. 224–226. [CrossRef]
34. Viñals, I.; Gimeno, P.; Ortega, A.; Miguel, A.; Lleida, E. In-domain Adaptation Solutions for the RTVE 2018 Diarization Challenge. In Proceedings of the IberSPEECH, Barcelona, Spain, 21–23 November 2018; pp. 220–223. [CrossRef]
35. Ghahabi, O.; Fischer, V. EML Submission to Albayzin 2018 Speaker Diarization Challenge. In Proceedings of the IberSPEECH, Barcelona, Spain, 21–23 November 2018; pp. 216–219. [CrossRef]
36. Ramos-Muguerza, E.; Docío-Fernández, L.; Alba-Castro, J.L. The GTM-UVIGO System for Audiovisual Diarization. In Proceedings of the IberSPEECH, Barcelona, Spain, 21–23 November 2018; pp. 204–207. [CrossRef]
37. India Massana, M.A.; Sagastiberri, I.; Palau, P.; Sayrol, E.; Morros, J.R.; Hernando, J. UPC Multimodal Speaker Diarization System for the 2018 Albayzin Challenge. In Proceedings of the IberSPEECH, Barcelona, Spain, 21–23 November 2018; pp. 199–203. [CrossRef]
38. Maurice, B.; Bredin, H.; Yin, R.; Patino, J.; Delgado, H.; Barras, C.; Evans, N.; Guinaudeau, C. ODESSA/PLUMCOT at Albayzin Multimodal Diarization Challenge 2018. In Proceedings of the IberSPEECH, Barcelona, Spain, 21–23 November 2018; pp. 194–198. [CrossRef]