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

Eduardo Lleida, Alfonso Ortega, Antonio Miguel, Virginia Bazán-Gil, Carmen Pérez, Manuel Gómez, Alberto de Prada

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[2], Carmen Pérez [2], Manuel Gómez [2] and Alberto de Prada [2]**

[1] Vivolab, Aragon Institute for Engineering Resarch (I3A), University of Zaragoza, Spain; http://www.vivolab.es
[2] Corporación Radiotelevisión Española, Spain; http:www.rtve.es
* Correspondence: lleida@unizar.es; University of Zaragoza, María de Luna 1, 50018 Zaragoza, Spain

Version October 18, 2019 submitted to Appl. Sci.

**Abstract:** Following the success of previous Albayzin evaluations supported by the Spanish Thematic Network on Speech Technologies (RTTH) since 2006, this paper describes the IberSpeech-RTVE Challenge at IberSpeech 2018, a new series of Albayzin evaluations focused on speech to text transcription, speaker diarization and multimodal diarization of TV broadcast content. For this purpose, the public Spanish television, *Corporación Radiotelevisión Española* (RTVE) and the RTVE Chair at the University of Zaragoza released a database with audiovisual and text documents. The database comprises different programs broadcast by RTVE from 2015 to 2018 covering a great variety of scenarios from studio to live broadcast, from read to spontaneous speech, different Spanish accents, including Latin-American accents and a great variety of contents. 18 international teams presented results in one or more evaluation task. This paper describes the database, 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) [1], in 2018 the broadcast media area was addressed. Jointly with Radio Televisión Española, RTVE [2], the Spanish Public Broadcast Corporation, Vivolab [3], the speech research group at the University of Zaragoza, proposed a set of technological evaluations in the areas of speech, speaker and face recognition. These evaluations were also supported by the RTVE Chair at the University of Zaragoza [4]. The dataset provided to participants included more than 500 hours of broadcast data, spanning a broad range of genres, subtitles, and human-revised transcriptions and speaker labels of part of the media contents.

Since 1996, when DARPA [1] presented the HUB-4 broadcast news corpora, several evaluations for

[1] Red Temática en Tecnologías del Habla (RTTH) http://www.rthabla.es
[2] Radiotelevisión Española (RTVE):http://www.rtve.es
[3] Vivolab http://vivolab.unizar.es
[4] Cátedra RTVE de la Universidad de Zaragoza: http://catedrartve.unizar.es

broadcast speech related tasks have been organized, most of them in English [2]. Other campaigns have been also carried out in other languages such as the ESTER evaluations in French [3,4], the TC-STAR evaluation in Mandarin [5], the 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 MGB Challenge with data in English and Arabic [12–14]. In areas different than 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 (S2T), speaker diarization (SD) and multimodal diarization (MD). 22 teams registered to the challenge and 18 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 and present their results along with 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

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 titles, duration and content of the shows included in the RTVE2018 database. The database is composed of 569 hours and 22 minutes of audio. About 460 hours are provided with the subtitles and about 109 hours 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.[5] The corpus is divided into 4 partitions, a *train* one, two development partitions *dev1*, *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 train partition consists of all the audio files without human-revised transcriptions, which means that only subtitles are available. The train partition can be used for any evaluation task. For development, two partitions are defined. Partition *dev1* contains about 53 hours of audios and their corresponding human-revised transcriptions. *Dev1* partition can be used for either development or training of the speech to text systems. Partition *dev2* contains about 15 hours of audios, human-revised transcriptions, speaker changes and their corresponding speaker labels. Additionally, *dev2* contains a 2 hour show annotated for multimodal diarization (face and speaker) and enrollment files (pictures, videos and audios) needed for speaker and face identification. Table 2 shows detailed information about the shows included in the development partitions.

RTVE2018 database includes a *test* partition with all the files needed to evaluate systems for speech to text, speaker and multimodal diarization. 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, *test* partition contains the enrollment files for the multimodal diarization challenge. It consists on 10 pictures and a 20 second video of the 39 characters to be identified.

[5] The respeaker 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.

Further detailed information about the RTVE2018 database content and formats can be found in the RTVE2018 database description report[6]. The RTVE2018 database is freely available subject to the terms of a licence agreement with the RTVE[7].

## 3. IberSpeech-RTVE 2018 Evaluation tasks

This section presents a brief summary of the three evaluations tasks. A more detailed description of the evaluation plans can be found in the Interspeech2018 web page[8] or Cátedra RTVE-UZ web page[9].

### *3.1. Speech-to-Text Challenge*

#### 3.1.1. Challenge Description and Databases

The Speech to Text transcription evaluation consisted of automatically transcribe different types of TV shows. The main objective was to evaluate the state of the art in automatic speech recognition for the Spanish language in the broadcast sector.

Training and Development data

The train partition consisted of all the audio files without human-revised transcriptions, which means that only subtitles were available. The train partition contains up to 460 hours of audio, half of them corresponding with a live magazine ("La Mañana"). Participants were free to use these audios as they considered appropiate.

For development, two partitions were defined. Partition *dev1* contains about 53 hours of audios and their corresponding human-revised transcriptions and partition *dev2* with about 15 hours of audios and their corresponding human-revised transcriptions and speaking-turns 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 training and development dataset 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 RTVE2018 training and development set or any other data to train their systems provided that these data were fully documented in the systems 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 8 different TV shows covering a variety of scenarios with a total of 39:07 hours of audio (see Table 3). The selected shows were different from those included in

[6] http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf
[7] http://catedrartve.unizar.es/rtvedatabase.html
[8] http://iberspeech2018.talp.cat/index.php/albayzin-evaluation-challenges/
[9] http://catedrartve.unizar.es/reto2018/evaluations2018.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 they do enjoy. |
| Al filo de lo imposible | 11:09:57 | This show broadcasts documentaries about mountaineering, climbing and other outdoor risk sports. It is a documentary series in which emotion, adventure sports and risk predominate. |
| Arranca en Verde | 05:38:05 | Contest dedicated to road safety presented. In it, viewers are presented with questions related to road safety in order to disseminate in a pleasant way the rules of the road and thus raise awareness about civic driving and respect for the environment. |
| Asuntos públicos | 69:38:00 | All the analysis of the news of the day and the live broadcast of the most outstanding information events. |
| Comando actualidad | 17:03:41 | A show that presents a current topic through the choral gaze of several street reporters. Four journalists who travel to the place where the news occurs, show them as they are and bring their personal perspective to the subject. |
| Dicho y Hecho | 10:06:00 | Game show in which a group of 6 comedians and celebrities compete against each other through hilarious challenges. |
| España en comunidad | 13:02:59 | Show that offers in-depth reports and current information about the different Spanish autonomous communities. It is made by the territorial and production centers of RTVE. |
| La mañana | 227:47:00 | Live Magazine, with a varied offer of contents for the whole family and with clear vocation of public service. |
| La tarde en 24H Economia | 04:10:54 | Program about economy |
| La tarde en 24H Tertulia | 26:42:00 | Talk show of political and economic news. (4/5 people) |
| La tarde en 24H Entrevista | 04:54:03 | In-depth interview with personalities from different fields. |
| La tarde en 24H El tiempo | 02:20:12 | Weather information of Spain, Europe and America. |
| Latinoamérica en 24H | 16:19:00 | Analysis and information show focused on Ibero-America, in collaboration with the Information Services of the International Area and the network of correspondents of RTVE. |
| Millennium | 19:08:35 | Debate show of ideas that pretends to be useful to the spectators of today, accompanying them in the analysis of everyday events. |
| Saber y Ganar | 29:00:10 | Daily contest presented that aims to disseminate culture in an entertaining way. Three contestants demonstrate their knowledge and mental agility, through a set of general questions. |
| La noche en 24H | 33:11:06 | Talk show with the best analysts to understand what has happened throughout the day. It contains interviews with some of the protagonists of the day. |
| **Total duration** | **569:22:04** | |

**Table 2.** Development dataset partition with shows and duration. (S2T: Speech to Text, 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. (S2T: Speech to Text, SD: Speaker Diarization, MD: Multimodal Diarization)

| 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** |

the development partition with human-revised transcriptions. Table 4 shows the main characteristics of the selected shows.

#### 3.1.2. Performance Measurement

S2T system output was evaluated with different metrics but they were ranked by the word error rate. All the participants had to provide, as S2T output for evaluation, a free-form text with no page, paragraphs, sentence or speaker breaks using the utf-8 charset per test file. The text might include punctuation marks to be evaluated with an alternative metric.

Primary metric

**Word Error Rate (WER)** was the primary metric for the S2T transcription task. The text was normalized removing all the punctuation marks, numbers were written with letters and text were lower-cased. The WER is defined as

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

where $N_r$ is the total 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 taking into account for the challenge ranking.

**Punctuation marks evaluation (PWER)** - The WER was computed with the punctuation marks given by the S2T transcription system. Periods and commas were processed as words.

**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 | A lot of speech overlap and speech inflections. About 8 speakers, most of them comedian. 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 | Lots of 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 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** | |

**Text Normalized Word Error Rate (TNWER)** - Text normalization techniques as stopword removal and lemmatization were applied to the S2T output. In this sense, common errors as verbal conjugations, gender or number substitutions, articles, determiners, and quantifiers deletion/insertions had not impact on the S2T performance evaluation metric. The same text normalization was applied to both the reference and automatic transcriptions before proceeding to calculate WER. The Freeling[10] 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 which originate 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 scoring metric as defined in the RT evaluations organized by NIST. An 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.** For training and development purposes, RTVE2018 database contained one training partition and a two development partitions. Around sixteen hours with diarization and

[10] http://nlp.lsi.upc.edu/freeling/

reference speech segmentation were included in the dev2 partition and might be used for any purpose including system development or training. The development data corresponded to two different debate shows, four episodes (7:26 hours) of *La noche en 24H*[11], where a group of political analysts comments what has happened throughout the day, and eight episodes (7:42 hours) of *Millenium*[12] where a group of experts debates about a current issue.

**Aragón Radio database.** The database donated by the *Corporación Aragonesa de Radio y Televisión*[13] (CARTV) consisted of around twenty hours of the Aragón Radio broadcast. This data set contains around 85% of speech, 62% of music and 30% of noise in 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[14] proposed for the 2010 Albayzin Audio Segmentation Evaluation [18,19] was recorded by the TALP Research Center from the UPC in 2009 under the Tecnoparla project[15] funded by the Generalitat de Catalunya. The *Corporació Catalana de Mitjans Audiovisuals*[16] (CCMA), owner of the multimedia content, allows its use for technology research and development. The database consists of around 87 hours of recordings in which speech can be found in a 92% of the segments, music is present a 20% of the time and noise in the background a 40%. Another class called *others* is defined which can be found a 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 contitions

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 audios 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 data set, as long as they were public access for all (not necessarily free).

Evaluation data

The evaluation of the diarization systems were done exclusively using the evaluation partition of the RTVE2018 database. This partition consisted of 22:45 hours 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 a 17,83%, 1 episode of "La Mañana" which represents 7.19% of the total and 9 episodes of "La Tarde en 24H" which represent 39.50% of the total audio. No a priori knowledge were 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[17], 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).

---

[11] http://www.rtve.es/alacarta/videos/la-noche-en-24-horas/
[12] http://www.rtve.es/alacarta/videos/millennium/
[13] http://www.cartv.es/
[14] http://www.ccma.cat/324/
[15] http://rua.ua.es/dspace/handle/10045/8626
[16] http://www.ccma.cat
[17] https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation

Given the dataset to evaluate $\Omega$, 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

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

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.

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

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. 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 diarizaton 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 segments of the same speaker with a silent of less that 2 seconds were merged and considered as a single segment. A forgiveness collar of 0.25 s., before and after each reference boundary, were considered in order to take into account both inconsistent human annotations and the uncertainty about when a speaker 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 which originate from the same speaker and face. For this evaluation, a list of characters to recognize were given. The rest of characters on the audiovisual document were discarded for the 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 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 had 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 use visual and audio information separately.

Development and evaluation data

- For development, *dev2* partition contained a 2 hour show "La noche en 24H" labeled with speaker and face timestamps. Enrollment files for the main characters were also provided. Enrollment

files consisted on pictures and short videos with the character speaking. Addtionally, the *dev2* partition contained around 14 hours of speaker diarization timestamps. No restrictions were placed on the use of any data outside the RTVE2018.
- For the evaluation, 3 television programs were distributed, one from "La Mañana" and two from "La Tarde en 24H Tertulia", which totals 4 hours. For enrollment, photos (10) and video (20 seconds) 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 anount 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 includes the time that is assigned to the wrong speaker or face, missed speech or face time and false alarm speech or face time:

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

Consecutive segments of the same speaker with a silent of less that 2 seconds were merged together and considered as a single segment. A forgiveness collar of 0.25 s., before and after each reference boundary, were considered in order to take into account both inconsistent human annotations and the uncertainty about when a speaker/face begins or ends.

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

$$DER_{total} = 0.5DER_{spk} + 0.5DER_{face} \quad (4)$$

## 4. Submitted systems

### *4.1. Speech to text challenge*

A total of 20 different systems from 7 participating teams were submitted. All of them presented results in the open-set condition and 3 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 S2T systems.

#### 4.1.1. Open-set condition systems

- **G1 - GTM-UVIGO** [20]. Multimedia Technologies Group, Universidad de Vigo, Spain.
  G1-GTM-UVIGO submitted two systems using as recognition engine the Kaldi toolkit[18]. Primary and contrastive systems differed in the language model used in the rescoring stage. The primary system used the 4-gram LM and the contrastive system used the RNNLM provided in Kaldi distribution. The acoustic models were trained using 109 hours of speech. 79 hours in Spanish (2006 TC-START[19]) and 30 in Galician (news database of Galicia, Transcrigal[20]). 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 has 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 has 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** [21]. 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 [21], which was built using the 3-fold augmented SAVAS[22], Albayzin[23], and Multext[24] corpora for 28 epochs. For this challenge, it was evolved for 2 new epochs using the same corpora in addition to the 3-fold augmented nearly perfectly aligned corpus obtained from the RTVE2018 dataset. A total of 897 hours were used for training. The language model was a 5-gram trained with the text data from the open-set dataset. The first constrastive system was also an evolution of the already existing E2E model, but in this case, it was evolved for one epoch using the 3-fold augmented corpora used in the primary system and a new Youtube RTVE corpus. The duration of the total amount of training audios was 1488 hours. The language model was a 5-gram trained with the text data from the open-set dataset. The second contrastive system was composed by a bidirectional LSTM-HMM acoustic model combined with a 3-gram language model for decoding and a 9-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** [22]. 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 submmited only a primary system. The recognition engine was an evolution of RETURNN[25] and RASR[26] and it was based on a hybrid LSTM-HMM acoustic model. Acoustic modeling was done using a bi-directional LSTM network with four layers and 512 LSTM units in each layer. 3800 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

[18] http://kaldi-asr.org/
[19] http://tcstar.org/
[20] http://metashare.elda.org/repository/browse/transcrigal-db/72ee3974cbec11e181b50030482ab95203851f1f95e64c00b842977a318ef641/
[21] https://arxiv.org/abs/1512.02595
[22] https://cordis.europa.eu/project/rcn/103572/factsheet/en
[23] http://catalogue.elra.info/en-us/repository/browse/ELRA-S0089/
[24] http://catalogue.elra.info/en-us/repository/browse/ELRA-S0060/
[25] https://github.com/rwth-i6/returnn
[26] https://www-i6.informatik.rwth-aachen.de/rwth-asr/

the single-pass HMM decoding was a 5-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 were used. Neither speaker nor domain adaptation nor model tuning were used

- **G14 - SIGMA** [23]. 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 a 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 hours from RTVE2018 (350 hours of manual transcription), VESLIM[27] (103 hours) and OWNMEDIA (162 hours of television programs). RTVE2018 database texts, news between 2015 and 2018, interviews and subtitles were used to train the language model.
- **G21 - EMPHATIC** [24]. SPIN-Speech Interactive Research Group, Universidad del País Vasco, Spain and Intelligent Voice, UK
  G21-EMPHATIC ASR system was based on the open-source Kaldi Toolkit. The Kaldi Aspire recipe[28] was used for building the DNNHMM acoustic model. Albayzin, Dihana[29], CORLEC-EHU[30] and TC-START databases with a total of 352 hours were used to train the acoustic models. Provided training and development audio files were subsampled to 8kHz before being used in the training and testing processes. A 3-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** [21]. 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 3-gram language model for decoding and a 9-gram language model for re-scoring lattices. The training and development set was aligned and filtered to get nearly 136 hours of audio with transcription. The acoustic model was trained with the nearly perfectly aligned partition, which was 3-fold augmented through the speed based augmentation technique. A total of 396 hours and 33 minutes were therefore used for training. The language models were estimated with the in-domain texts compiled from the RTVE2018 dataset. The first constrastive system was set up using the same configuration of the primary system, but the acoustic model was estimated using the 3-fold augmented acoustic data of the perfectly aligned partition. A total of 258 hours and 27 minutes were employed for training. The same data were used to build the the second contrastive system but it was an E2E recognition system which follows the architecture used for the open-set condition. The language model was a 5-gram.
- **G7 - MLLP-RWTH** [22]. MLLP, Machine Learning and Language Processing, Universidad Politécnica de Valencia, Spain and Human Language Technology and Pattern Recognition, RWTH Aachen University, Germany.

---

27 https://ieeexplore.ieee.org/document/1255449/
28 https://github.com/kaldi-asr/kaldi/tree/master/egs/aspire
29 http://universal.elra.info/product_info.php?products_id=1421
30 http://gtts.ehu.es/gtts/NT/fulltext/RodriguezEtal03a.pdf

G7-MLLP-RWTH submitted two systems. The recognition engine was the TLK toolkit decoder[25]. The ASR consisted of a BLSTM-HMM acoustic model, and a combination of both RNN and TV-show adapted n-gram language models. The training and development set were aligned and filtered to get nearly 218 hours of audio with transcription. For the primary system, all aligned data from train, dev1 and dev2 partitions, 218 hours, were used for acoustic model training. For the contrastive system, only a reliable set of 205 hours of the train and dev1 partitions were user for training the acoustic models. The language models were estimated with the in-domain texts compiled from the RTVE2018 dataset with a lexicon of 132K words.

- **G14 - SIGMA** [23]. Sigma AI, Madrid, Spain.
  G14-SIGMA submitted only a primary system. The system had the same architecture that the one submitted for the open-set conditions, but only 350 hours 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 9 participating teams were submitted. 6 of them presented results in the closed-set condition and 5 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[31] was used with Kaldi and data from VoxCeleb1 and VoxCeleb2 databases[32]. This DNN maps variable length speech segments into fixed dimension vectors called xvectors. The strategy followed for diarization consisted of three main stages. First stage of extraction of xvector and grouping using the clustering algorithm "Chinese Whispers". In the second stage, each of the audio segments were processed to extract one or more xvectors using a sliding window of 10 seconds with half a second of displacement between successive windows. These vectors were grouped using the clustering algorithm "Chinese Whispers" obtaining the clusters that define 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 were also trained with external data.
- **G11-ODESSA** [26]. EURECOM, LIMSI, CNRS, France.
  A primary system resulting from the combination at similarity matrix level of 3 systems, one trained according to the closed-set condition and 2 trained with two external databases (NIST SRE and Voxceleb) were submitted. The first contrastive system used MFCC features, 1-second uniform segmentation, x-vector representation trained on NIST SRE data and AHC clustering. The second contrastive system was the same that the one for the closed-set, but where the training data was replaced with the Voxceleb data.
- **G20-STAR-LAB** [27]. STAR Lab, SRI International, USA.
  The training signals were extracted from the databases NIST SRE[33] 2004-2008, NIST SRE 2012, Mixer6[34], Voxceleb1, and Voxceleb2. Data augmentation was used to degrade signals. STAR-Lab used the embeddings and diarization system developed for the speaker recognition competition NIST 2018 [28]. It incorporated modifications in the detection of voice activity and in the

[31] http://kaldi-asr.org/models.html
[32] http://www.robots.ox.ac.uk/~vgg/data/voxceleb/
[33] https://www.nist.gov/itl/iad/mig/speaker-recognition
[34] https://catalog.ldc.upenn.edu/LDC2013S03

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** [29]. SPIN-Speech Interactive Research Group, Universidad del País Vasco, Spain, Intelligent Voice, UK.
  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 system used MFCC parameters, the representation of speaker embeddings was obtained through an end-to-end model using convolutional (CNN) and recurrent (LSTM) networks.
- **G22-JHU** [30]. 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 4 different types of embeddings extractors: x-vector-basic, x-vector-factored, i-vector-basic and BNF-i-vector were used. All the systems followed the structure: parameter extraction, voice activity detector, embeddings extraction, PLDA scoring, fusion and grouping of speakers.

#### 4.2.2. Closed-set condition systems

- **G4-VG** [31]. Voice Group, Advanced Technologies Application Center-CENATAV, Cuba.
  The submitted systems used a classic structure based on BIC segmentation, hierarchical agglomerative grouping and re-segmentation by hidden Markov models. The toolbox S4D[35] was used. The difference between the submitted systems were the feature extraction method ranging from classic ones as MFCC, LFCC and LPCC to new ones such as Mean Hilbert Envelope Coefficients [32], Medium Duration Modulation Coefficients and Power Normalization Cepstral Coefficients [33].
- **G8-AUDIAS-UAM** [34]. 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 Long Short-Term Memory (BLSTM) 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** [35]. 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 represents 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 did not.
- **G11-ODESSA** [26]. EURECOM, LIMSI, CNRS, France.
  The primary system was the fusion at diarization hypothesys level of the two contrastive systems. The first contrastive system used ICMC features (constant Q, Mel-frequency cepstral coefficients), 1-second uniform segmentation, BK representation and AHC clustering, while the second one used MFCC features, BiLSTM based speaker change detection, triplet-loss neural embedding representation and affinity propagation clustering.
- **G19-EML** [36]. European Media Laboratory GmbH, Germany.
  The submitted diarization system was designed primarily for on-line applications. Every 2

[35] https://projets-lium.univ-lemans.fr/s4d/

seconds, it made a decision about the identity of the speaker without using future information. It used speaker vectors based on the transformation of a GMM supervector.

- **G22-JHU** [30]. 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 4 participating teams were submitted.

#### 4.3.1. System descriptions

- **G1-GTM-UVIGO** [37]. 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 finally the results of each system were fused to better adjust the speech of speakers.
- **G2-GPS-UPC** [38]. 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 and his face did not appear, and faces that appeared but did not speak.
- **G9-PLUMCOT** [39].LIMSI, CNRS, France. The submitted systems (primary and 2 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** [39]. 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 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 were submitted by 7 teams, 12 systems to the open-set condition and 6 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 is 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

**Table 5.** Speech to text open-set condition. WER(%) per TV show and team. Team and system descriptions in section 4.1.1. See table 4 for show descriptions. (P: Primary, C#: contrastive #)

| **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 |

text resources used to train the systems, except for the commercial systems whose information was not provided, the G1-MLLP-RHTW team was the one that used the most resources for the training acoustic and language models, followed by the G14-SIGMA and G6-VICOMTECH-PRHLT group. The correlation between performance and training resources is clear. Also, it should be taken into account that the G14-SIGMA acoustic models were trained using 350 hours of the training set with manual transcription whereas G6-VICOMTECH-PRHLT used aligned and filtered data from the training and development set. Regarding the WER per TV shows, the variance across shows is quite high. The WER per show varied from 7.43 – 34.45% for the most accurate system. LA24H (Latinoamérica en 24H) and EC (España en Comunidad) were the ones with lower WER. A priori, it was not expect the good results obtained by LA24H as it contains Spanish accents from Latin-America. However most of the time the show contains a high quality voice-over in terms of acoustic environment and pronunciation. The worst results were given by the "Dicho y Hecho" quiz show due to the acoustic environment and speech inflections of the participants. Also, the WER variability among episodes of the same show is high. Figure 1 shows a boxplot of the WER variability per show for the best system. Also, table 9 shows the overall text normalized WER (TNWER). The reference and test text are normalized in terms of removing stopwords and changing each word by its lemma. The relative WER reduction is 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 3.1.2.2) for the submitted systems. There are 4 systems submitted with dots and 3 with periods and commas. In all the cases, the performance is degraded mainly by the increase in the number of deletions and insertions related with periods and commas. Team G6 obtained much better results with the second contrastive system than the primary. Primary system used an E2E approach but the second contrastive system was a BLSTM-HMM system with n-gram language model. Training data was the same for primary and contrastive system.

#### 5.1.2. Closed-set condition results

Regarding closed-set condition results, table 8 presents the overall results of the three primary systems. Results are presented as in the open-set condition. The best results, with a WER of 19.57%, were obtained by the G14-SIGMA system. The G14-SIGMA system in the closed-set condition was the same that the one used in the open-set condition, the difference was 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 additional 265 hours of audio and text coming from news, interviews and subtitles. This data augmentation allowed to reduce 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 is how they obtained audio a text data from the train and development partitions. G14 used manual transcription of 350 hours of the train partition for both

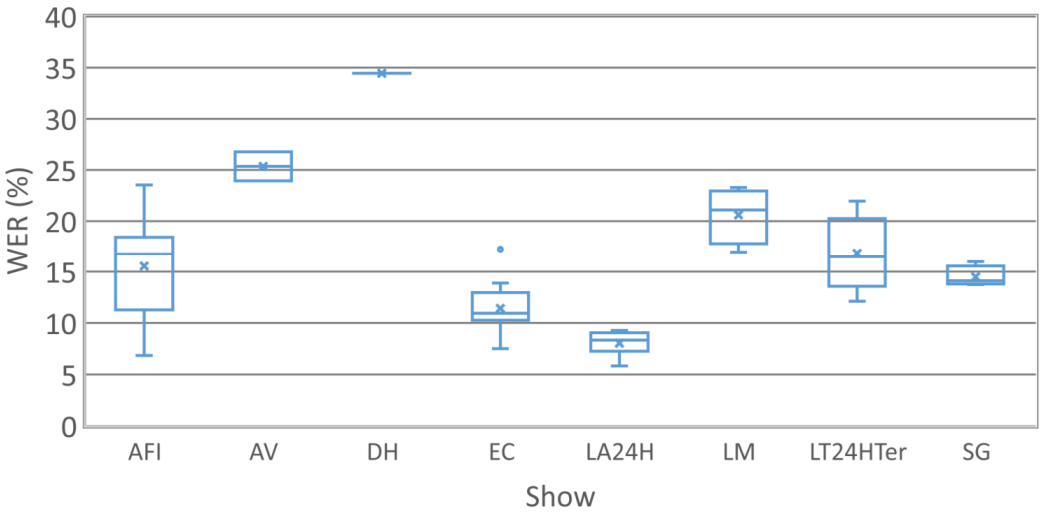


**Figure 1.** Boxplot with means (x) of the WER by shows for the best system in the open-set condition. See table 4 for show descriptions.

**Table 6.** Speech to text open-set condition. TNWER (%) for the primary systems per show and team. Team and system descriptions in section 4.1.1. See table 4 for show descriptions.

| **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 PWER (%) for the primary systems per team. Team and system descriptions in section 4.1.1. (P: Primary, C#: contrastive #)

| PUNC. MARKS | PERIODS | PERIODS | | | PERIODS & 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 |

**Table 8.** Speech to text closed-set condition. WER(%) per TV show and team. Team and system descriptions in section 4.1.2. See table 4 for show descriptions. (P: Primary, C#: contrastive #)

| **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** |
| **Overal 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 in section 4.1.2.

| **TEAM** | G6 | G7 | G14 |
|---|---|---|---|
| **Overall WER** | 22.22 | 21.98 | **19.57** |
| **Overall TNWER** | 20.71 | 19.75 | **17.90** |

acoustic and language models. However G6 and G7 used automatic aligned and filtered audios from the train and development partitions, a total of 136 hours in G6 and 218 hours in G7, for training acoustic models and all text files given in RTVE2018 database for training language models. Results correlate quite well with the amount of training data. G14 got the best results and G7 obtained slightly better results than G6. In terms of the TNWER it is interesting to note that G14 almost got the same TNWER (17.90%) than in the open-set condition (17.26%) which means that most of the errors in the closed-set conditions were related with stopwords, verbal conjugations and word number or gender. Comparison between open-set and closed-set conditions for G6 and G7 systems were not possible as they used different ASR architectures.

### *5.2. Speaker diarization evaluation*

The speaker diarization challenge achieved a participation of 9 teams submitting a total of 26 systems, 13 for each condition, open-set and closed-set. Results are presented for the open-set and closed-set conditions.

#### 5.2.1. Open-set condition results

5 teams participated in the open-set condition evaluation. 4 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 have included on this review as their results are quiet impressive. Table 10 presents results for each submitted system by team. G1-GTM-UVIGO got a DER of 11.4% which is almost half of next system, the contrastive C1 system from G11 team with a DER of 20.3%. The most significant difference between the G1 systems and the rest of systems is the low speaker error, G1 got 6.6% and the next got 16.8%. Figure 2 presents the estimation of the average number of speaker of the best systems of each team. The average number of speaker was 27. G1-GTM-UVIGO system made a close estimation of the number of speakers with an average of 28. However, the rest of the systems underestimated by a big difference. Clearly, estimation of the number of speakers is a key point to explain the better results of the G1-GTM-UVIGO with respect to others. In terms of the shows, see table 11, the lower DER was obtained by "La Tarde en 24H Tertulia" which is a talk show of political and economic news with an average of 20 speakers with a good quality audio and almost no speech

**Table 10.** Open-set condition Speaker Diarization: DER (%), Missed Speech (%), False Speech (%) and Speaker Error (%) per team. Team and system descriptions in section 4.2.1. (P: Primary, C#: contrastive #)

| **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 in section 4.2.1. (P: Primary, C#: contrastive #)

| **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 |

overlap. The rest of the shows include studio and exterior shots with a higher number of speakers, 65 for LM (La Mañana), 34 for LA24H (Latino América en 24H) and 29 for EC (España en Comunidad).

### 5.2.2. Closed-set condition results

6 teams participated in the closed-set condition evaluation. Table 12 shows results of all submitted systems. 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% in contrastive system 2. One of the most noticeable difference between the more accurate system and the rest is the speaker error rate which correlates with the average estimation of the number of speakers. Figure 2 shows the number of speaker estimation by primary systems per team. G10-VIVOLAB system gave very close estimates of the number of speakers and the rest clearly gave an underestimation. It is interesting to compare result of G11-ODESSA between open and closed conditions. C2 system was the same for both conditions, the only difference was the training dataset, Voxceleb data for the open-set condition and RTVE2018 dataset for the closed-set condition. Results were quite similar with a slight improvement on the open-set condition. Regarding the results per show, table 13 presents diarization errors per show for the best system of each team.

**Table 12.** Closed-set condition Speaker Diarization: DER (%), Missed Speech (%), False Speech (%) and Speaker Error (%) per team. Team and system descriptions in section 4.2.2. (P: Primary, C#: contrastive #)

| **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 |

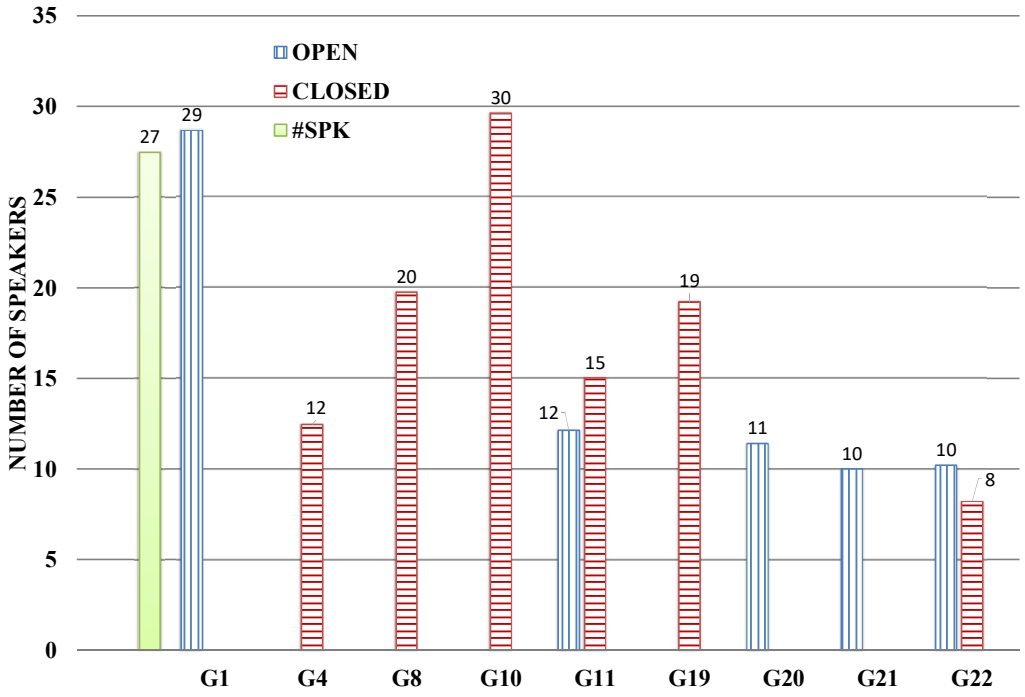


**Figure 2.** Estimation of the average number of speaker by the primary speaker diarization system of each team. #SPK is the real average number of speaker. Team and system descriptions in section 4.2.1

**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 in section 4.2.2. (P: Primary, C#: contrastive #)

| **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 |

**Table 14.** Multimodal Diarization: DER (%), Missed Speech/Fase (%), False Speaker/Face (%) and Speaker/Face Error (%) for the best systems. Team and system descriptions in section 4.3. (P: Primary, C#: contrastive #)

| **TEAM** | G1 | | G2 | | G9 | | G11 | |
|---|---|---|---|---|---|---|---|---|
| **MODALITY/SYSTEM** | FACE/P | SPKR/P | FACE/P | SPKR/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 |

### *5.3. Multimodal diarization evaluation*

Table 14 presents results with the best combinations of face and speaker systems per team and TV shows. Note that the number of characters to be identify were 39. Three from four systems gave a very similar DER in the margin of 23% to 30%. Best result was given by the system presented by G9-PLUMCOT with 23.39% of DER. It is remarkable that thanks to the merger of both modalities, G9-PLUMCOT team managed to reduce from 29.08% of the G11-ODESSA system (independent modalities) to 23.39%. Both teams were using the same systems for each modality. Note that the results of the G2-GPS-UPC team are significantly worse due to a poor adjustment of the speaker diarization system. This fact can be seen in results segregated by modality in table 14. Except the G2-GPS-UPC system, the speech modality gave significantly better results than the face modality. G1-GTM-UVIGO speaker diarization primary system was the one that provides the best results 17.3% followed by G9-PLUMCOT system with 17.6%. For the face modality, the system used by G9-PLUMCOT and G11-ODESSA gave the best results followed but the G2-GPS-UPC system. In terms of the TV shows, "La Mañana" (LM-20170103) presented the highest difficulty for both face and speech modality mainly due to the fact that it is a live magazine with lots of speaker overlaps and exterior shots. However, "La tarde en 24 hora, Tertulia" (LT24HTer-*) has a few overlapped speech, up to 5 speakers and characters per show, most of them in studio shots.

## 6. Conclusions

The IberSpeech-RTVE 2018 challenge was a new Albayzin evaluation focused on speech to text transcription, speaker diarization and multimodal diarization of TV broadcast content. We achieved wide international participation from 18 teams with 5 teams participated in more than one evaluation. A new dataset, named RTVE2018 database, was released containing more that 500 hours of TV shows.

The speech to text challenge achieved a participation of 7 teams submitting a total of 18 different systems, 12 for the open-set condition and 6 for the closed-set condition. The evaluation was done over 8 different shows covering a wide range of acoustic conditions. The speech to text performance showed a high variability across the shows with a WER per show from 7.43% to 34.45% for the best system in the open-set condition. The most accurate system gave an overall WER of 16.56%. Closed-set conditions got worse results due to the difficulty of getting clean transcriptions from the training data set. The best result was obtained by training the acoustic models with 350 hours of manual transcriptions of the training set with a WER of 19.57%. Text normalization deleting stopwords and lematizing gave a small improvement in the WER.

The speaker diarization challenge achieved a participation of 9 teams submitting a total of 26 systems, 13 for each condition, open-set and closed-set. The evaluation was done over 4 different shows covering a wide range of acoustic conditions, number of speakers and overlap speech. The best systems in open-set and closed-set conditions gave a DER of 11.4% and 17.3% respectively. There was a big gap

in terms of DER between the best systems and the rest in both conditions. Both systems gave a very close estimation of the average number of speakers, however the rest made a clear underestimation.

The multimodal diarization challenge achieved a participation of 4 teams. The evaluation was done over 2 shows, one episode of a live magazine (La Mañana) and 2 episodes of talk show of political and economic news. The task was to identify speakers and faces of a closed set of 39 characters. The most accurate system gave a multimodal DER of 23.4%. In terms of separate face and speaker diarization, the best system gave a speaker DER of 17.6% and a face DER of 29.2%. In the three episodes the speaker DER was much lower than the face DER.

We plan to continue with the Albayzin evaluations in the next IberSpeech conference in 2020. An extension of the database with new annotated audiovisual material for training, development and testing audiovisual technologies is being prepared.

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

**Funding:** This work has been supported by the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the project TIN2017-85854-C4-1-R, Government of Aragón (Reference Group T36_17R) and co-nanced 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 2019 organizers.

**Conflicts of Interest:** The authors declare no conict of interest.

## References

1. Garofolo, J.; Fiscus, J.; Fisher, W. Design and preparation of the 1996 HUB-4 broadcast news benchmark test corpora. *Proc. DARPA Speech Recognitnio Workshop* **1997**.
2. Graff, D. An overview of Broadcast News corpora. *Speech Communication* **2002**, *37*, 15 – 26. doi:https://doi.org/10.1016/S0167-6393(01)00057-7.
3. Galliano, S.; Geoffrois, E.; Gravier, G.; f. 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, 2006, pp. 315–320.
4. Galliano, S.; Gravier, G.; Chaubard, L. The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. Interspeech; , 2009; pp. 2583–2586.
5. D., M.; O., H.; K., C. Evaluation of Automatic Speech Recognition and Spoken Language Translation within TC-STAR: results from the first evaluation campaign. Proceedings of LREC'06, Genoa, Italy., 2006.
6. Butko, T.; Nadeu, C. Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion. *EURASIP Journal on Audio, Speech, and Music Processing* **2011**, p. 1. doi:10.1186/1687-4722-2011-1.
7. Zelenák, M.; Schulz, H.; Hernando, J. Speaker diarization of broadcast news in Albayzin 2010 evaluation campaign. *EURASIP Journal on Audio, Speech, and Music Processing* **2012**, p. 19. doi:10.1186/1687-4722-2012-19.
8. Ortega, A.; Castan, D.; Miguel, A.; Lleida, E. The Albayzin-2012 audio segmentation evaluation. Iberspeech, 2012.
9. Ortega, A.; Castan, D.; Miguel, A.; Lleida, E. The Albayzin-2014 audio segmentation evaluation. Iberspeech, 2014.
10. Castán, D.; Ortega, A.; Miguel, A.; Lleida, E. Audio segmentation-by-classification approach based on factor analysis in broadcast news domain. *EURASIP Journal on Audio, Speech, and Music Processing* **2014**, p. 34. doi:10.1186/s13636-014-0034-5.
11. Ortega, A.; Viñals, I.; Miguel, A.; Lleida, E. The Albayzin-2016 Speaker Diarization Evaluation. Iberspeech, 2016.
12. Bell, P.; Gales, M.J.F.; Hain, T.; Kilgour, J.; Lanchantin, P.; Liu, X.; McParland, A.; Renals, S.; Saz, O.; Wester, M.; Woodland, P.C. The MGB challenge: Evaluating multi-genre broadcast media recognition. 2015

IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015, Scottsdale, AZ, USA, December 13-17, 2015, pp. 687–693. doi:10.1109/ASRU.2015.7404863.
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. 2016 IEEE Spoken Language Technology Workshop, SLT, San Diego, CA, USA, December 13-16, 2016, pp. 279–284. doi:10.1109/SLT.2016.7846277.
14. Ali, A.; Vogel, S.; Renals, S. Speech recognition challenge in the wild: Arabic MGB-3. 2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017, Okinawa, Japan, December 16-20, 2017, pp. 316–322. doi:10.1109/ASRU.2017.8268952.
15. Versteegh, M.; Thiollière, R.; Schatz, T.; Cao, X.; Anguera, X.; Jansen, A.; Dupoux, E. The zero resource speech challenge 2015. INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 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. 2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017, Okinawa, Japan, December 16-20, 2017, pp. 323–330. doi:10.1109/ASRU.2017.8268953.
17. Eskevich, M.; Aly, R.; Racca, D.N.; Ordelman, R.; Chen, S.; Jones, G.J.F. The Search and Hyperlinking Task at MediaEval 2014. Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Catalunya, Spain, October 16-17, 2014.
18. Zelenak, M.; Schulz, H.; Hernando, J. Albayzin 2010 Evaluation Campaign: Speaker Diarization. VI Jornadas en Tecnologías del Habla, FALA 2010. Vigo, 2010.
19. Zelenak, M.; Schulz, H.; Hernando, J. Speaker diarization of broadcast news in Albayzin 2010 evaluation campaign. EURASIP Journal on Audio, Speech, and Music Processing, 2012.
20. Docío-Fernández, L.; García-Mateo, C. The GTM-UVIGO System for Albayzin 2018 Speech-to-Text Evaluation. Proc. IberSPEECH, 2018, pp. 277–280. doi:10.21437/IberSPEECH.2018-58.
21. 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. Proc. IberSPEECH, 2018, pp. 267–271. doi:10.21437/IberSPEECH.2018-56.
22. 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. Proc. IberSPEECH, 2018, pp. 257–261. doi:10.21437/IberSPEECH.2018-54.
23. 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. Proc. IberSPEECH, 2018, pp. 262–266. doi:10.21437/IberSPEECH.2018-55.
24. Dugan, N.; Glackin, C.; Chollet, G.; Cannings, N. Intelligent Voice ASR system for Iberspeech 2018 Speech to Text Transcription Challenge. Proc. IberSPEECH, 2018, pp. 272–276. doi:10.21437/IberSPEECH.2018-57.
25. del Agua, M.; Giménez, A.; Serrano, N.; Andrés-Ferrer, J.; Civera, J.; Sanchis, A.; Juan, A. The translectures-UPV toolkit. Advances in Speech and Language Technologies for Iberian Languages, 2014, pp. 269–278.
26. Patino, J.; Delgado, H.; Yin, R.; Bredin, H.; Barras, C.; Evans, N. ODESSA at Albayzin Speaker Diarization Challenge 2018. Proc. IberSPEECH, 2018, pp. 211–215. doi:10.21437/IberSPEECH.2018-43.
27. Castan, D.; McLaren, M.; Nandwana, M.K. The SRI International STAR-LAB System Description for IberSPEECH-RTVE 2018 Speaker Diarization Challenge. Proc. IberSPEECH, 2018, pp. 208–210. doi:10.21437/IberSPEECH.2018-42.
28. McLaren, M.; Ferrer, L.; Castan, D.; Nandwana, M.; Travadi, R. The sri-con-usc nist 2018 sre system description. NIST 2018 Speaker Recognition Evaluation, 2018.
29. Khosravani, A.; Glackin, C.; Dugan, N.; Chollet, G.; Cannings, N. The Intelligent Voice System for the IberSPEECH-RTVE 2018 Speaker Diarization Challenge. Proc. IberSPEECH, 2018, pp. 231–235. doi:10.21437/IberSPEECH.2018-48.
30. Huang, Z.; García-Perera, L.P.; Villalba, J.; Povey, D.; Dehak, N. JHU Diarization System Description. Proc. IberSPEECH, 2018, pp. 236–239. doi:10.21437/IberSPEECH.2018-49.
31. Campbell, E.L.; Hernandez, G.; Calvo de Lara, J.R. CENATAV Voice-Group Systems for Albayzin 2018 Speaker Diarization Evaluation Campaign. Proc. IberSPEECH, 2018, pp. 227–230. doi:10.21437/IberSPEECH.2018-47.

32. Sadjadi, S.O.; Hansen, J.H. Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification. *Speech Communication* **2015**, *72*, 138 – 148. doi:https://doi.org/10.1016/j.specom.2015.04.005.
33. Kim, C.; Stern, R.M. Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2016**, *24*, 1315–1329. doi:10.1109/TASLP.2016.2545928.
34. Lozano-Diez, A.; Labrador, B.; de Benito, D.; Ramirez, P.; T. Toledano, D. DNN-based Embeddings for Speaker Diarization in the AuDIaS-UAM System for the Albayzin 2018 IberSPEECH-RTVE Evaluation. Proc. IberSPEECH, 2018, pp. 224–226. doi:10.21437/IberSPEECH.2018-46.
35. Viñals, I.; Gimeno, P.; Ortega, A.; Miguel, A.; Lleida, E. In-domain Adaptation Solutions for the RTVE 2018 Diarization Challenge. Proc. IberSPEECH, 2018, pp. 220–223. doi:10.21437/IberSPEECH.2018-45.
36. Ghahabi, O.; Fischer, V. EML Submission to Albayzin 2018 Speaker Diarization Challenge. Proc. IberSPEECH, 2018, pp. 216–219. doi:10.21437/IberSPEECH.2018-44.
37. Ramos-Muguerza, E.; Docío-Fernández, L.; Alba-Castro, J.L. The GTM-UVIGO System for Audiovisual Diarization. Proc. IberSPEECH, 2018, pp. 204–207. doi:10.21437/IberSPEECH.2018-41.
38. 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. Proc. IberSPEECH, 2018, pp. 199–203. doi:10.21437/IberSPEECH.2018-40.
39. Maurice, B.; Bredin, H.; Yin, R.; Patino, J.; Delgado, H.; Barras, C.; Evans, N.; Guinaudeau, C. ODESSA/PLUMCOT at Albayzin Multimodal Diarization Challenge 2018. Proc. IberSPEECH, 2018, pp. 194–198. doi:10.21437/IberSPEECH.2018-39.