The ALBAYZIN 2024 Search on Speech Evaluation Plan

Javier Tejedor[1] and Doroteo T. Toledano[2]

[1] Escuela Politécnica Superior, Universidad San Pablo-CEU, CEU Universities, Madrid, Spain.
javiertejedornoguerales@gmail.com

[2] AUDIAS - Audio, Data Intelligence and Speech, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Madrid, Spain

Abstract. This document presents the evaluation plan for the coming ALBAYZIN 2024 Search on Speech evaluation. This evaluation aims to find a list of terms/queries in audio archives and is divided into two different tasks: Spoken Term Detection and Query-by-Example Spoken Term Detection. Spoken Term Detection employs a list of terms for searching whereas Query-by-Example Spoken Term Detection makes use of acoustic examples for searching. The Spoken Term Detection task cannot make use of prior knowledge of the list of terms when processing the audio. On the other hand, Query-by-Example Spoken Term Detection cannot make use of prior knowledge of the correct word/phone transcription of the acoustic examples when conducting the search. In this evaluation, participants are encouraged to build end-to-end systems for both tasks.

1 Introduction

The ALBAYZIN 2024 Search on Speech evaluation is supported by the Spanish Thematic Network on Speech Technology (RTTH)[3] and is organized by Universidad San Pablo-CEU and AUDIAS from Universidad Autónoma de Madrid. The evaluation workshop will be part of *IberSpeech 2024* to be held in Aveiro (Portugal), in November 2024.

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

2 Evaluation description

The Search on Speech evaluation consists of two different tasks:

– **Spoken Term Detection (STD)**, where the input to the system is a list of terms, but these terms are unknown when processing the audio. This task

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

must generate a set of occurrences for each term detected in the audio files, along with their timestamps and score as output. This is the same task as in NIST STD 2006 evaluation [1] and Open Keyword Search in 2013 [3], 2014 [4], 2015 [5], and 2016 [6].
- **Query-by-Example Spoken Term Detection (QbE STD)**, where the input to the system is an acoustic query and hence a prior knowledge of the correct word/phone transcription corresponding to each query cannot be used. This task must generate a set of occurrences for each query detected in the audio files, along with their timestamps and score as output, as in the STD task. This QbE STD is the same task as those proposed in MediaEval 2011, 2012, and 2013 [2].

A baseline system for the STD task that participants can employ for their system construction can be found at *https://github.com/javiertejedornoguerales/Whisper_STD*.

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

In case a Large Vocabulary Continuous Speech Recognition (LVCSR) system is employed to construct the system for Spoken Term Detection and/or Query-by-Example Spoken Term Detection tasks, the way in which the LVCSR dictionary has been built **must** be fully described in the system description paper.

2.1 Primary and contrastive systems

Participants could submit their system/s either for the Spoken Term Detection task, Query-by-Example Spoken Term Detection task or for both tasks. Participants are required to submit one primary system and up to 4 contrastive systems for any task. Both development and test output files must be submitted by participants. In this way, overfitting and calibration issues can be detected and the robustness of the proposed methodology can be evaluated and compared to other methodologies. Participants will be ranked in these tasks according to the performance attained by their primary systems on the test data.

3 Database description

Three different databases will be employed in this evaluation for the STD and QbE STD tasks: (1) MAVIR database, which has been used in previous ALBAYZIN Search on Speech evaluations, and comprises a set of talks extracted from the Spanish MAVIR workshops[4] held in 2006, 2007 and 2008 (Corpus MAVIR 2006, 2007 and 2008) corresponding to Spanish language, (2) RTVE

[4] http://www.mavir.net

database, which comprises different Radio TeleVisión Española (RTVE) programs recorded from 1960 to 2023, and (3) COSER database, which consists of several interviews with elderly people from rural areas covering different dialectal varieties in Spain. For all databases, three separate datasets (i.e., for training, development, and test) will be provided to participants.

3.1 Training data

Training data provided by the evaluation organizers belong to MAVIR, RTVE and COSER databases. However, we do not limit the amount of training data that can be employed to build the systems and hence any kind of data can be used for system training provided that these data are fully documented in the system description paper. In addition, these training data can be used for participants as they consider more suitable (i.e., training, development, etc.).

Regarding the MAVIR database, about 4 hours of speech extracted from 5 audio files will be provided as training material. The speech data were originally recorded in several audio formats (PCM mono and stereo, MP3, etc). All data were converted to PCM, 16khz, single channel, and 16 bits per sample WAV files using the Sox tool [5]. The corresponding word transcription of the speech material will also be provided.

Regarding the RTVE database, about 900 hours of speech extracted from different TV shows will be provided as training material. The speech data are provided in AAC format, stereo, 44100hz, and variable bit rate. The speech files can be easily converted to PCM, 16khz, single channel, and 16 bits per sample WAV files with the *ffpmeg* tool[6] and the following command:

ffmpeg -i <fich>.aac -ar 16000 –ac 1 <fich>.wav,

where *<fich>* is the name of the audio file in the original AAC format. From all the training data, about 460 hours of speech are provided with subtitles, though these could not contain an accurate word transcription, and the rest of the speech data are provided with human-revised word transcriptions. The RTVE data are available to the evaluation participants subject to the terms of license agreement with the RTVE. The license agreement can be downloaded from Cátedra RTVE-UZ web page (http://catedrartve.unizar.es/rtvedatabase.html). Participants must sign the agreement and send a scanned copy attached to the email. A digital signature is also valid. A copy signed by RTVE representative will be returned following the instructions given in the web page. RTVE authorizes the use of the contents released for the evaluation, for their use in research works, to all those participants. The authorization will be valid for three years from the date of the public communication of the results of the evaluation. After this period, if necessary, an extension may be requested for the same use.

[5] http://sox.sourceforge.net/
[6] https://ffmpeg.org/

Regarding the COSER database, around 60 hours of speech from different interviews will be provided as training material. The speech data were originally recorded in several audio formats (e.g., mono and stereo) and different sampling frequency. All the speech data were converted to PCM, 16khz, single channel and 16 bits per sample WAV files using the Sox tool. The corresponding word transcription of the speech material will also be provided, along with turn-level timestamps.

3.2 Development data

Development data belong to the same databases for which training data are provided (i.e., Spanish MAVIR workshop material, RTVE programs and COSER interviews). However, we do not limit the amount of development data that can be employed to tune the system parameters and hence any kind of data can be used for system tuning provided that these data are fully documented in the system description.

Participants must submit an output result file for the development data corresponding to the MAVIR database, a different output result file for the development dataset named *dev2* corresponding to the RTVE database, and an output result file for the development data corresponding to the COSER database, even if participants employ more development data for system tuning.

For the Spoken Term Detection task, orthographic transcriptions of the selected list of terms along with the occurrences and timestamps for each of these terms will be provided at due time.

For the Query-by-example Spoken Term Detection task, audio files with the queries and all the occurrences and timestamps for each query will also be provided at due time.

For the MAVIR database, the development speech data amount to about 1 hour of speech material in total, extracted from 2 audio files. For the Spoken Term Detection task, the development list of terms consists of about 375 different terms whose length ranges from 5 to 27 single graphemes. A term can be composed by one or more words. About 100 queries, with three examples per query, will be extracted from the Spoken Term Detection development list of terms to compose the Query-by-Example Spoken Term Detection development query list. These three examples consist of an example extracted from the MAVIR development speech data, and two other examples recorded by evaluation organizers. These two examples amount to 3 seconds of speech and have been recorded at a sampling frequency of 16Khz in WAV format and mono with the microphone of an HP ProBook Core i5, 7th Gen and with a Sennheiser SC630 USR CTRL microphone with noise cancellation, respectively. These MAVIR development data can be used as participants find more suitable (i.e., for training, development, etc.).

For the RTVE database, two different datasets will be provided as development data: *dev1* and *dev2*. The *dev1* RTVE development speech dataset amounts to more than 50 hours of speech material in total, extracted from 5 TV shows, for which human-revised word transcriptions will be provided. The

dataset named *dev2*, for which participants have to submit their detection results on this dataset for RTVE, amounts to about 12 hours of speech material in total, extracted from 10 audio files, for which human-revised word transcriptions will also be provided. For this *dev2* dataset, the list of terms consists of about 400 different terms whose length ranges from 4 to 25 single graphemes for Spoken Term Detection task. A term can be composed by one or more words. About 100 queries, with three examples per query, will be extracted from the Spoken Term Detection development list of terms to compose the Query-by-Example Spoken Term Detection development query list. These three examples consist of an example extracted from the RTVE *dev2* development speech data, and two other examples recorded by evaluation organizers. These two examples amount to 3 seconds of speech and have been recorded at a sampling frequency of 16Khz in WAV format and mono with the microphone of an HP ProBook Core i5, 7th Gen and with a Sennheiser SC630 USR CTRL microphone with noise cancellation, respectively. These RTVE development data can be used as participants find more suitable (i.e., for training, development, etc.).

For the COSER database, the development speech data amount to about 2 hours in total, extracted from different interviews. For the Spoken Term Detection task, the development list of terms consists of about 200 different terms whose length ranges from 4 to 23 single graphemes. A term can be composed by one or more words. For the Query-by-Example Spoken Term Detection task, about 100 queries, with three examples per query, extracted from the Spoken Term Detection test list of terms, will be used. These three examples consist of an example extracted from the COSER development speech data, and two other examples recorded by evaluation organizers. These two examples amount to 3 seconds of speech and have been recorded at a sampling frequency of 16Khz in WAV format and mono with the microphone of an HP ProBook Core i5, 7th Gen and with a Sennheiser SC630 USR CTRL microphone with noise cancellation, respectively. These COSER development data can be used as participants find more suitable (i.e., for training, development, etc.).

3.3 Test data

Three databases will be employed for system evaluation: MAVIR, RTVE and COSER.

The MAVIR data are built from the MAVIR material explained before. The test speech data amount to about 2 hours in total, extracted from 3 audio files. For the Spoken Term Detection task, only the list of terms used for evaluation will be provided. This list consists of about 200 different terms whose length ranges from 4 to 28 single graphemes. A term can be composed by one or more words. For the Query-by-Example Spoken Term Detection task, about 100 queries, with three examples per query, extracted from the Spoken Term Detection test list of terms, will be used for evaluation. These three examples consist of an example extracted from the MAVIR test speech data, and two other examples recorded by evaluation organizers. These two examples amount to 3 seconds of speech and have been recorded at a sampling frequency of 16Khz in

WAV format and mono with the microphone of an HP ProBook Core i5, 7th Gen and with a Sennheiser SC630 USR CTRL microphone with noise cancellation, respectively.

The RTVE data consist of the RTVE program material explained before. The test speech data amount to about 2 hours in total, extracted from different TV shows. For the Spoken Term Detection task, only the list of terms used for evaluation will be provided. This list consists of about 200 different terms whose length ranges from 4 to 27 single graphemes. A term can be composed by one or more words. For the Query-by-Example Spoken Term Detection task, about 100 queries, with three examples per query, extracted from the Spoken Term Detection test list of terms, will be used for evaluation. These three examples consist of an example extracted from the RTVE test speech data, and two other examples recorded by evaluation organizers. These two examples amount to 3 seconds of speech and have been recorded at a sampling frequency of 16Khz in WAV format and mono with the microphone of an HP ProBook Core i5, 7th Gen and with a Sennheiser SC630 USR CTRL microphone with noise cancellation, respectively.

The COSER data consist of a small subset of the interviews explained before. The test speech data amount to about 2 hours in total, extracted from different interviews. For the Spoken Term Detection task, only the list of terms used for evaluation will be provided. This list consists of about 200 different terms whose length ranges from 3 to 26 single graphemes. A term can be composed by one or more words. For the Query-by-Example Spoken Term Detection task, about 100 queries, with three examples per query, extracted from the Spoken Term Detection test list of terms, will be used for evaluation. These three examples consist of an example extracted from the COSER test speech data, and two other examples recorded by evaluation organizers. These two examples amount to 3 seconds of speech and have been recorded at a sampling frequency of 16Khz in WAV format and mono with the microphone of an HP ProBook Core i5, 7th Gen and with a Sennheiser SC630 USR CTRL microphone with noise cancellation, respectively.

None of these test data are allowed to be used as training or development data in any form.

4 Evaluation of system performance

The Actual Term Weighted Value (ATWV) [1] will be the primary metric for the STD and QbE STD tasks. Participants will be ranked for each database individually from the ATWV obtained on the test data. Therefore, participants are not compelled to submit detection results for all the databases. Maximum Term Weighted Value (MTWV) scores and Detection Error Tradeoff (DET) [1] curves for STD and QbE STD tasks will also be computed.

5 General evaluation conditions

5.1 Data organization

The datasets will be available through a web page (MAVIR and COSER data) and ftp (RTVE data); instructions for downloading will be given to participants at due time for the release of training and development data.

Training data. For the Spoken Term Detection and Query-by-Example Spoken Term Detection tasks, the training data, which correspond to the MAVIR, RTVE and COSER databases, will consist of the following elements:

- *MAVIR/audio* - a folder with the MAVIR training audio files.
- *MAVIR/transcription* - a folder with the word transcription (sentence-level timestamps will be provided) of the MAVIR training audio files.
- *RTVE2018DB/train/audio* - a folder with some RTVE training audio files.
- *RTVE2018DB/train/srt* - a folder with the subtitles of the RTVE training audio files in *RTVE2018DB/train/audio*. These may not contain accurate word transcriptions.
- *RTVE2020DB/test/audio/S2T* - a folder with some RTVE training audio files.
- *RTVE2020DB/test/references/S2T/stm* - a folder with the human-revised word transcription of the RTVE training audio files in *RTVE2020DB/test/audio/S2T*.
- *RTVE2022DB/train/audio* - a folder with some RTVE training audio files.
- *RTVE2022DB/train/stm* - a folder with the human-revised word transcription of the RTVE training audio files in *RTVE2022DB/train/audio*.
- *RTVE2022DB/test/audio/S2T* - a folder with some RTVE training audio files.
- *RTVE2022DB/test/references/S2T* - a folder with the human-revised word transcription of the RTVE training audio files in *RTVE2022DB/test/audio/S2T*.
- *COSER/audio* - a folder with the COSER training audio files.
- *COSER/transcription* - a folder with the word transcription (turn-level timestamps will be provided) of the COSER training audio files.
- Check the README file for additional data information.

Development data. For the Spoken Term Detection and Query-by-Example Spoken Term Detection tasks, the development data that correspond to the MAVIR, RTVE and COSER databases will consist of the following elements:

- *MAVIR/audio* - a folder with the MAVIR development audio files.
- *MAVIR/occurrences* - a folder with the word transcription and timestamps of all the occurrences of the selected list of terms/queries corresponding to the MAVIR development data.

- *MAVIR/queries* - a folder with the three acoustic examples per query (in folders query1, query2 and query3) that serve as development queries for the Query-by-Example Spoken Term Detection task for the MAVIR database. The folder query1 contains the acoustic example extracted from the MAVIR development speech data. The folder query2 contains the acoustic example recorded with the microphone of an HP ProBook Core i5, 7th Gen. The folder query3 contains the acoustic example recorded with a Sennheiser SC630 USR CTRL microphone with noise cancellation.
- *RTVE2018DB/dev1/audio* - a folder with the RTVE development *dev1* dataset audio files.
- *RTVE2018DB/dev1/trn* - a folder with the RTVE development *dev1* dataset human-revised word transcriptions.
- *RTVE2018DB/dev2/audio* - a folder with the RTVE development *dev2* dataset audio files, on which participants have to submit detection results for RTVE development data, along with some additional audio files.
- *RTVE2018DB/dev2/trn* - a folder with the RTVE development *dev2* dataset human-revised word transcriptions.
- *RTVE2018DB/dev2/occurrences* - a folder with the word transcription and timestamps of all the occurrences of the selected list of terms/queries corresponding to the RTVE development *dev2* dataset.
- *RTVE2018DB/dev2/queries* - a folder with the three acoustic examples per query (in folders query1, query2 and query3) that serve as development queries for the Query-by-Example Spoken Term Detection task for the RTVE database. The folder query1 contains the acoustic example extracted from the RTVE *dev2* development speech data. The folder query2 contains the acoustic example recorded with the microphone of an HP ProBook Core i5, 7th Gen. The folder query3 contains the acoustic example recorded with a Sennheiser SC630 USR CTRL microphone with noise cancellation.
- *COSER/audio* - a folder with the COSER development audio files.
- *COSER/occurrences* - a folder with the word transcription and timestamps of all the occurrences of the selected list of terms/queries corresponding to the COSER development data.
- *COSER/queries* - a folder with the three acoustic examples per query (in folders query1, query2 and query3) that serve as development queries for the Query-by-Example Spoken Term Detection task for the COSER database. The folder query1 contains the acoustic example extracted from the COSER development speech data. The folder query2 contains the acoustic example recorded with the microphone of an HP ProBook Core i5, 7th Gen. The folder query3 contains the acoustic example recorded with a Sennheiser SC630 USR CTRL microphone with noise cancellation.
- *scoring* - a folder with the scoring scripts along with the necessary input files for evaluating the systems for Spoken Term Detection and Query-by-Example Spoken Term Detection tasks. Both tasks will be scored using the NIST STD scoring tool [1].
- *doc* - a folder with relevant evaluation information: example output file, evaluation plan, data organization, README file, etc.

Test data. For the Spoken Term Detection and Query-by-Example Spoken Term Detection tasks, the test data that correspond to MAVIR, RTVE and COSER databases will consist of the following elements:

- *MAVIR/test/audio/SoS* - a folder with the test audio files for the MAVIR database.
- *MAVIR/data* - a folder with the text files that contain the list of test terms/queries for the MAVIR database.
- *RTVE2024/test/audio/SoS* - a folder with the test audio files for the RTVE database.
- *RTVE2024/data* - a folder with the text files that contain the list of test terms/queries for the RTVE database.
- *COSER/test/audio/SoS* - a folder with the test audio files for the COSER database.
- *COSER/data* - a folder with the text files that contain the list of test terms/queries for the COSER database.
- *MAVIR/queries* - a folder with the three acoustic examples per query (in folders query1, query2 and query3) that serve as test queries for the Query-by-Example Spoken Term Detection task for the MAVIR database. The folder query1 contains the acoustic example extracted from the MAVIR test speech data. The folder query2 contains the acoustic example recorded with the microphone of an HP ProBook Core i5, 7th Gen. The folder query3 contains the acoustic example recorded with a Sennheiser SC630 USR CTRL microphone with noise cancellation.
- *RTVE2024/queries* - a folder with the three acoustic examples per query (in folders query1, query2 and query3) that serve as test queries for the Query-by-Example Spoken Term Detection task for the RTVE database. The folder query1 contains the acoustic example extracted from the RTVE test speech data. The folder query2 contains the acoustic example recorded with the microphone of an HP ProBook Core i5, 7th Gen. The folder query3 contains the acoustic example recorded with a Sennheiser SC630 USR CTRL microphone with noise cancellation.
- *COSER/queries* - a folder with the three acoustic examples per query (in folders query1, query2 and query3) that serve as test queries for the Query-by-Example Spoken Term Detection task for the COSER database. The folder query1 contains the acoustic example extracted from the COSER test speech data. The folder query2 contains the acoustic example recorded with the microphone of an HP ProBook Core i5, 7th Gen. The folder query3 contains the acoustic example recorded with a Sennheiser SC630 USR CTRL microphone with noise cancellation.
- *scoring* - a folder with the scoring script, and the necessary input files for evaluating the systems for MAVIR, RTVE and COSER databases, except for the ground-truth (.rttm) files, which will be released once the results are officially published.
- *doc* - a folder with relevant evaluation information: example output file, evaluation plan, data organization, README file, etc.

5.2 System output format

Detection results must be sent in a single file according to the 'stdlist' XML format specified in the NIST STD 2006 evaluation plan [1] both for the Spoken Term Detection and Query-by-Example Spoken Term Detection tasks. An example of the output format, the necessary input files, and the NIST STD scoring tool will be provided with the training and development data. Please note that for these tasks, timestamps for each detection are relevant, since they are taken into account by the NIST STD scoring tool to evaluate if each term detection is correct or not. Higher scores mean more confidence in the detection appearing in the corresponding speech file between the given timestamps.

5.3 Submissions

Registration rules. Interested groups must register for the evaluation before July 31st, 2024 in the following web page: http://catedrartve.unizar.es/albayzin2024.html.

Submission procedure. Recognition results for Spoken Term Detection and/or Query-by-Example Spoken Term Detection tasks, along with the corresponding file describing the system or systems and the obtained results (on, at least, development data), must be submitted by participants. Instructions for output file and system description paper submissions will be announced to the registered participants at due time for the release of evaluation data.

Filenames must be constructed according to the following pattern:

<Group>_<Task>_<SysID>_<Set>_<Data>.xml

where *<Group>* is the acronym of the group according to the registration data, *<Task>* is STD or QbESTD for Spoken Term Detection and Query-by-Example Spoken Term Detection tasks respectively, and *<SysID>* is a code that identificates the system as primary (pri) or contrastive (con1, con2, etc). The *<Set>* field must be set to DEV for development data and EVAL for test data, and the *<Data>* field is MAVIR, RTVE, or COSER to identify the corresponding database. As an example, if the group HLPGA builds a primary system for both tasks, one contrastive system for the Spoken Term Detection task, and two contrastive systems for the Query-by-Example Spoken Term Detection task for all the databases, the following files must be submitted:

HLPGA_STD_pri_DEV_MAVIR.xml
HLPGA_STD_pri_DEV_RTVE.xml
HLPGA_STD_pri_EVAL_MAVIR.xml
HLPGA_STD_pri_EVAL_COSER.xml
HLPGA_STD_pri_EVAL_RTVE.xml
HLPGA_STD_con1_DEV_MAVIR.xml
HLPGA_STD_con1_DEV_RTVE.xml
HLPGA_STD_con1_EVAL_MAVIR.xml

HLPGA_STD_con1_EVAL_COSER.xml
HLPGA_STD_con1_EVAL_RTVE.xml
HLPGA_QbESTD_pri_DEV_MAVIR.xml
HLPGA_QbESTD_pri_DEV_RTVE.xml
HLPGA_QbESTD_pri_EVAL_MAVIR.xml
HLPGA_QbESTD_pri_EVAL_COSER.xml
HLPGA_QbESTD_pri_EVAL_RTVE.xml
HLPGA_QbESTD_con1_DEV_MAVIR.xml
HLPGA_QbESTD_con1_DEV_RTVE.xml
HLPGA_QbESTD_con1_EVAL_MAVIR.xml
HLPGA_QbESTD_con1_EVAL_COSER.xml
HLPGA_QbESTD_con1_EVAL_RTVE.xml
HLPGA_QbESTD_con2_DEV_MAVIR.xml
HLPGA_QbESTD_con2_DEV_RTVE.xml
HLPGA_QbESTD_con2_EVAL_MAVIR.xml
HLPGA_QbESTD_con2_EVAL_COSER.xml
HLPGA_QbESTD_con2_EVAL_RTVE.xml

Note that field values (e.g., acronym of the group) should not contain underscores ('_'), so as not to confuse the parsing.

System description. Research groups must provide a file with the description of the submitted systems. If multiple systems are submitted for a particular task, the description must explicitly designate one of them as the primary system, the remaining ones being contrastive systems. The system description paper should give the readers a good sense of what the system is about, keeping in mind the following guidelines:

- Write for your audience. Remember that the reader is not you but other system developers who may not be familiar with your technique/algorithm. Clearly explain your method so they can understand what you did.
- A superficial description would leave other system developers clueless of what you did. Be as complete as possible, but not to the extent of including pseudo-code. Include all the relevant information, in such a way that other groups can build the system on their own.
- Include references to techniques, algorithms, subsystems, etc., used by your systems but not described in detail in the document.
- Avoid jargon and abbreviations without any prior context.
- Include the results obtained (at least) on the development data.

Participants can choose between two submission ways:

- The first way relies on editing the system description paper following the IberSpeech 2024 paper submission template so that the submitted paper (describing the system/s and the results) will appear in the IberSpeech 2024 proceedings. Moreover, participants will also have the chance to submit an

extended version of this paper to a journal. This submission way implies sending one or more representatives to the evaluation workshop, to be held in Aveiro, Portugal as part of IberSpeech 2024 (November 2024).
- The second way demands a free-format document in which participants describe the submitted system/s along with the results, but this will not appear in the IberSpeech 2024 proceedings. In this case, participants are allowed to present on-line their system/s without physically attending the conference, or send a video to the evaluation organizers explaining their submitted system/s, which will be shown during the evaluation workshop.

In any case, the system description paper should, at least, include the following sections:

1 Introduction

2 System A (name of the submitted system)

2.1 System description
Clearly describe the methods and algorithms used in system A.

2.2 Train and development data
Describe all the data and/or systems directly or indirectly used in developing system A, including the source, acquisition conditions, size, publishing year and any other pertinent information.

3 System B (name of another submitted system)
This section is similar to section 2 but for another system. If system B is a contrastive system, note the differences from the primary system. A new section should be added for each submitted system.

4 Results
Results of the submitted systems on (at least) the development data provided by the organizers.

5 References
List of papers relevant to the techniques, algorithms, data, etc. used by the submitted systems.

5.4 Schedule

- May 20th, 2024. Registration opens.
- June 3rd, 2024. Release of the training and development data.
- July 31st, 2024. Registration deadline.
- September 2nd, 2024. Release of the evaluation data. System submission opens.
- October 18th, 2024 (23:59, GMT +1). System submission deadline.

- October 31st, 2024. Results and ground-truth files are distributed to the participants.
- November 12th, 2024.Official results are presented publicly and published.
- November 12th, 2024. Iberspeech 2024 Albayzin Evaluations special session in Aveiro.

6 Additional information for participants and summary of evaluation rules

- Interested groups must register for the evaluation before July 31st, 2024 in the following web page: http://catedrartve.unizar.es/albayzin2024.html.
- Starting from June 3rd, 2024, and once registration data are validated, the training and development data will be released only to registered participants.
- The test data will be released by September 2nd, 2024. Recognition results must be submitted to the organizing team by the established deadline: October 18th, 2024 (23:59 GMT+1). The paper submission deadline is October 30th, 2022 (23:59 GMT+1).
- Registered groups commit themselves to use the provided data only for research purposes, distribution being allowed only with explicit permission of the ALBAYZIN 2024 Search on Speech Evaluation organizing team. Registered participants are allowed to use MAVIR data to develop or evaluate their own systems, provided that they acknowledge that use by means of the following references:
  "**MAVIR corpus**:
  http://www.lllf.uam.es/ESP/CorpusMavir.html",
  and that they cite the Albayzin 2024 Search on Speech system description paper that will be included in the IberSpeech 2024 Proceedings. Regarding the RTVE data, RTVE will allow the use of these data to the participants only for research purposes in case they request them. This request will be valid for three years starting from the date of the public communication of the results of the evaluation. Once this 3-year agreement finishes, participants can make a new request to continue using them. Please, refer to the corresponding license agreement for full information and details. In case a discrepancy between this and the license is found, the license prevails.
  Authors are also required to cite the Albayzin 2024 Search on Speech system description paper that will be included in the IberSpeech 2024 Proceedings in case of using the RTVE and/or COSER data.
- No manual intervention is allowed for each system developed to generate the final output file and hence, all the developed systems must be fully automatic. Listening to the test data, or any other human interaction with the test data is forbidden before all the results have been submitted.
- In case the participant site has submitted a paper to appear in the IberSpeech proceedings, it is mandatory to send one or more representatives to the evaluation workshop, to be held in Aveiro, Portugal as part of IberSpeech 2024

(November 2024). However, in case the participant site has just submitted a free-format document with the system description+results, it is allowed to present on-line their system without physically attending the conference, or send a video to the evaluation organizers explaining their submitted systems, which will be shown during the evaluation workshop.
– This plan might be modified due to new restrictions or unplanned needs, to detected errors or inaccuracies. Updated versions of this plan, if any, will be announced through the Search on Speech evaluation website and emailed to the registered participants.

7 Acknowledgements

This work was partially supported by the project "Multi-task and Semi-supervised Deep Learning for Speech and Audio Processing" (PID2021-125943OB-I00) from the Spanish Ministry of Science, Innovation and Universities.

References

1. Fiscus, J.G., Ajot, J.G., Garofolo, J.S., Doddington, G.: Results of the 2006 spoken term detection evaluation. In: Proc. of ACM SIGIR. pp. 1–4 (2007)
2. Metze, F., Anguera, X., Barnard, E., Davel, M., Gravier, G.: Language independent search in mediaeval's spoken web search task. Computer Speech and Language (2014)
3. NIST: NIST Open Keyword Search 2013 Evaluation (OpenKWS13). National Institute of Standards and Technology (NIST), Washington DC, USA, 1 edn. (July 2013), http://www.nist.gov/itl/iad/mig/openkws13.cfm
4. NIST: NIST Open Keyword Search 2014 Evaluation (OpenKWS14). National Institute of Standards and Technology (NIST), Washington DC, USA, 1 edn. (July 2014), http://www.nist.gov/itl/iad/mig/openkws14.cfm
5. NIST: NIST Open Keyword Search 2015 Evaluation (OpenKWS15). National Institute of Standards and Technology (NIST), Washington DC, USA, 1 edn. (July 2015), http://www.nist.gov/itl/iad/mig/openkws15.cfm
6. NIST: NIST Open Keyword Search 2016 Evaluation (OpenKWS16). National Institute of Standards and Technology (NIST), Washington DC, USA, 1 edn. (July 2016), http://www.nist.gov/itl/iad/mig/openkws16.cfm