The ALBAYZIN 2016 Search on Speech Evaluation Plan

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

[1] FOCUS S.L., Madrid, Spain,
javiertejedornoguerales@gmail.com

[2] ATVS - Biometric Recognition Group,
Universidad Autónoma de Madrid, Spain

Abstract. This document presents the evaluation plan for the coming ALBAYZIN 2016 Search on Speech evaluation. This evaluation aims at finding 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.

1 Introduction

The ALBAYZIN 2016 Search on Speech evaluation is supported by the Spanish Thematic Network on Speech Technology (RTTH)[3] and is organized by FOCUS S.L. and ATVS - Biometric Recognition Group of Autonomous University of Madrid. The evaluation workshop will be part of *IberSpeech 2016* to be held in Lisbon, Portugal from 23 to 25 November 2016.[4]

This evaluation involves searching in audio content a list of terms/queries and hence it is suitable for groups working on speech indexing and retrieval and on speech recognition in general as well. In other words, this evaluation focuses on retrieving the audio files that contain any of those terms/queries. This year, the evaluation organizers will also provide word lattices generated from a speech recognizer. This tries to encourage research groups working outside speech recognition (e.g., text information retrieval) to take part in the evaluation.

2 Evaluation description

The Search on Speech evaluation consists of two different tasks:

[3] http://lorien.die.upm.es/∼lapiz/rtth/

[4] http://iberspeech2016.inesc-id.pt/

- **Spoken Term Detection (STD)**, where the input to the system is a list of terms, but this is unknown when processing the audio. This is the same task as in NIST STD 2006 evaluation [2] 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 example per query and hence a prior knowledge of the correct word/phone transcription corresponding to each query cannot be made. This task must generate a set of occurrences for each query detected in the audio files, along with their timestamps and scores as output, as in the STD task. This QbE STD is the same task as those proposed in MediaEval 2011, 2012, and 2013 [1].

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. This evaluation will define two different sets of terms/queries for STD and QbE STD tasks. One in-vocabulary (INV) set of terms/queries and one out-of-vocabulary (OOV) set of terms/queries. The OOV set of terms/queries will be composed by out-of-vocabulary words for the LVCSR system. This means that, in case participants employ an LVCSR system for processing the audio for any task (STD, QbE STD), these OOV terms (i.e., all the words that compose the term) must be previously removed from the system dictionary and hence, other methods (e.g., phone-based systems) have to be used for searching OOV terms/queries.

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 (presumably, the best one) and up to 2 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. Participants are allowed to use any available resources, as long as their use is documented in the system description paper.

3 Database description

Two different databases will be employed in this evaluation. MAVIR database, used in previous ALBAYZIN Search on Speech evaluations, and EPIC database. For MAVIR database, three separate datasets (i.e., for training, development, and test) will be provided to participants. For EPIC database, only test data will be provided. This will allow organizers to measure the generalization capability of the systems in an unseen domain.

3.1 Training data

Training data provided by the evaluation organizers belong to a set of talks extracted from the Spanish MAVIR workshops[5] held in 2006, 2007 and 2008 (Corpus MAVIR 2006, 2007 and 2008) corresponding to Spanish language. 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. 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, 16 bits per sample using the Sox tool [6]. The corresponding word transcription of the speech material will be also provided.

3.2 Development data

Development data will also belong to Spanish MAVIR workshop material. 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 paper. It must be noted that development data cannot be used for system training at all (in terms of acoustic models, language models, or in general, any kind of *training*). Participants must submit an output result file for the development data provided by the organizers, no matter participants employ more development data for system tuning. For Spoken Term Detection task, orthographic transcriptions of the selected list of terms along with the occurrences and timestamps for each of these terms corresponding to development data will be provided at due time in the evaluation web page. The development list of terms consists of about 375 different terms (some of these being INV terms and the rest are OOV) whose length ranges from 5 to 27 single graphemes for Spoken Term Detection task. A term can be composed by one or more words. About 100 queries (some of these being INV queries and the rest are OOV), with one example per query, will be extracted from the Spoken Term Detection development list of terms to compose the Query-by-Example Spoken Term Detection

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

[6] http://sox.sourceforge.net/

task input. Occurrences and timestamps of these queries will be also provided in the evaluation web page. Development data amount to about 1 hour of speech material in total, extracted from 2 audio files.

3.3 Test data

Two different datasets will be employed for system evaluation: MAVIR and EPIC.

MAVIR dataset is built from the MAVIR material explained before. The speech data amount to about 2 hours in total, extracted from 3 audio files. For Spoken Term Detection task, only the list of terms used for evaluation will be provided. This list consists of about 200 different terms (some of these being INV terms and the rest are OOV) 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 (some of these being INV queries and the rest are OOV), with one example per query, extracted from the Spoken Term Detection test list of terms, will be used for evaluation.

EPIC database comprises data from European Parliament speeches recorded in 2004 in English, Spanish and Italian, along with their corresponding simultaneous interpretations to the other languages. Only the Spanish original speeches will be used for evaluation. For Spoken Term Detection task, only the list of terms used for evaluation will be provided. This list consists of about 180 different terms (some of these being INV terms and the rest are OOV) whose length ranges from 6 to 16 single graphemes. A term is composed by a single word. For the Query-by-Example Spoken Term Detection task, about 100 queries (some of these being INV queries and the rest are OOV), with one example per query, extracted from the Spoken Term Detection test list of terms, will be used for evaluation.The EPIC[7] database is distributed by the European Language Resources Association (ELRA). It is free for research purposes as long as it is downloaded by FTP, but participants need to request the corpus directly to ELRA and sign the corresponding license with ELRA before downloading it. We will only provide the list of terms and the audio queries. The Spanish original speeches used for the evaluation are videos in MPEG format. For labelling and to extract the audio queries, the audios of those videos have been extracted and converted to single channel, 16 KHz, 16 bits with the following ffmpeg[8] command: ffmpeg -i "infile.mpg" -vn -ar 16000 -ac 1 "outfile.wav". We suggest using this command to extract the audios from the videos, but this is not mandatory. Given that the original test materials from EPIC are videos, participants are allowed to use the video materials along with the audio, although the test has been designed to use the audio only.

[7] "European Parliament Interpretation Corpus (EPIC), ELRA catalogue (http://catalog.elra.info), ISLRN: 716-168-855-843-2, ELRA ID: ELRA-S0323"

[8] ffmpeg version N-79068-g6b7ce0e (https://ffmpeg.org/)

4 Evaluation of system performance

The Actual Term Weighted Value (ATWV) [2] will be the primary metric for the STD and QbE STD tasks. DET curves for STD and QbE STD tasks will be also computed.

5 General evaluation conditions

5.1 Data organization

All the datasets (except the EPIC corpus) will be available through a web page; instructions for downloading will be given to participants at due time for the release of training and development data.

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

- *audio* - a folder with the training audio files.
- *transcription* - a folder with the word transcription (no timestamps will be given) of the training audio files.
- Check the README file for additional data information.

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

- *audio* - a folder with the development audio files.
- *transcription* - a folder with the word transcription and timestamps of all the occurrences of the selected list of terms/queries corresponding to development data.
- *lattices* - a folder with the word lattices generated by a speech recognizer when processing the development audio files.
- *queries* - a folder with the acoustic examples that serve as development queries for the Query-by-Example Spoken Term Detection task.
- *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 [2].
- *doc* - a folder with relevant evaluation information: example output file, evaluation plan, data organization, README file, etc.

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

- *data/MAVIR* - a folder with the test audio files and a text file that contains the list of test terms/queries.
- *data/EPIC* - a folder with a text file that contains the list of test terms/queries. It must be noted that the audio files have to be downloaded by participants according to the procedure explained in Section 3.3.
- *queries/MAVIR* - a folder with the acoustic examples that serve as test queries for the Query-by-Example Spoken Term Detection task for the MAVIR database.
- *queries/EPIC* - a folder with the acoustic examples that serve as test queries for the Query-by-Example Spoken Term Detection task for the EPIC database.
- *lattices/MAVIR* - a folder with the word lattices generated by a speech recognizer when processing the test audio files for the MAVIR database.
- *lattices/EPIC* - a folder with the word lattices generated by a speech recognizer when processing the test audio files for the EPIC database.
- *scoring* - a folder with the scoring script, and the necessary input files for evaluating the systems for MAVIR and EPIC 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 for Spoken Term Detection and Query-by-Example Spoken Term Detection tasks must be sent in a single file according the 'stdlist' XML format specified in the NIST STD 2006 evaluation plan [2]. An example of the output format, the necessary input files, and the NIST STD scoring tool will be provided with the 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 15th 2016, by contacting the organizing team at:

javiertejedornoguerales@gmail.com

with copy (cc) to Iberspeech 2016 Evaluation organizers at:

luisjavier.rodriguez@ehu.eus
lapiz@die.upm.es
alberto.abad@l2f.inesc-id.pt
ortega@unizar.es
ajst@ua.pt

and providing the following information:

– Research group (name and acronym).
– Institution (university, research center, etc).
– Contact Person (name).
– Email address.

Submission procedure. Recognition results for Spoken Term Detection and/or Query-by-Example Spoken Term Detection tasks, along with the corresponding PDF file describing the system/s, must be submitted by participants. Instructions for output files and system description paper submission 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 for MAVIR database and EPIC for EPIC database. As an example, if the group HLPGA builds a primary system for both tasks, one contrastive system for Spoken Term Detection task, and two contrastive systems for Query-by-Example Spoken Term Detection task, the following files must be sent:

HLPGA_STD_pri_DEV_MAVIR.xml
HLPGA_STD_pri_EVAL_MAVIR.xml
HLPGA_STD_pri_EVAL_EPIC.xml
HLPGA_STD_con1_DEV_MAVIR.xml
HLPGA_STD_con1_EVAL_MAVIR.xml
HLPGA_STD_con1_EVAL_EPIC.xml
HLPGA_QbESTD_pri_DEV_MAVIR.xml
HLPGA_QbESTD_pri_EVAL_MAVIR.xml
HLPGA_QbESTD_pri_EVAL_EPIC.xml
HLPGA_QbESTD_con1_DEV_MAVIR.xml
HLPGA_QbESTD_con1_EVAL_MAVIR.xml
HLPGA_QbESTD_con1_EVAL_EPIC.xml
HLPGA_QbESTD_con2_DEV_MAVIR.xml

HLPGA_QbESTD_con2_EVAL_MAVIR.xml
HLPGA_QbESTD_con2_EVAL_EPIC.xml

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

System description. Research groups must provide a PDF 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.

To keep formal homogeneity, it is **mandatory** to edit the system description paper by means of the IberSpeech 2016 paper submission template (Springer LNAI format), available at the following site: http://iberspeech2016.inesc-id.pt/. 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 References
List of papers relevant to the techniques, algorithms, data, etc. used by the submitted systems.

5.4 Schedule

- June 1, 2016. Registration opens.
- June 30, 2016. Release of the training and development data.
- July 15, 2016. Registration deadline.
- September 15, 2016. Release of the evaluation data. System submission opens.
- October 15, 2016 (24:00, GMT +1). Deadline for the submission of results and system description paper.
- October 31, 2016. Results and ground-truth files are distributed to the participants.
- November 23-25, 2016. IberSpeech 2016, Lisbon, Portugal: Evaluation results are presented and discussed.

6 Additional information for participants and summary of evaluation rules

- Interested groups must register for the evaluation before July 15th 2016, by contacting the organizing team at:

  javiertejedornoguerales@gmail.com

  with copy (cc) to Iberspeech 2016 Evaluation organizers at:

  luisjavier.rodriguez@ehu.eus
  lapiz@die.upm.es
  alberto.abad@l2f.inesc-id.pt
  ortega@unizar.es
  ajst@ua.pt

  and providing the following information:

  - Research group (name and acronym).
  - Institution (university, research center, etc.).
  - Contact person (name).
  - Email address.
- Starting from June 30th 2016, and once registration data are validated, the training and development data will be released via web (only to registered participants).
- The evaluation dataset will be released by September 15, 2016. Recognition results along with the system description paper must be submitted to the organizing team by the established deadline: October 15, 2016 until 24:00 GMT+1.

- Research groups must provide a description of the submitted systems, according to the guidelines given in Section 5.3. For the sake of formal homogeneity, **it is mandatory** to edit the system description paper by means of the Iberspeech 2016 paper submission template (Springer LNAI format) available at:
  http://www.springer.de/comp/lncs/authors.html
- Registered groups commit themselves to use the provided data only for research purposes, distribution being allowed only with explicit permission of the ALBAYZIN 2016 Search on Speech Evaluation organizing team. Registered participants are allowed to use the data to develop or evaluate their own systems, provided that they acknowledge that use by means of the following reference:
  "**MAVIR corpus**.
  http://www.lllf.uam.es/ESP/CorpusMavir.html"
  and that they cite the Albayzin 2016 Search on Speech system description paper that will be included in the Iberspeech 2016 Proceedings.
  The use of the EPIC data is regulated in the license that participants need to sign with ELRA to download the 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.
- Each participating site is required to send one or more representatives to the evaluation workshop, to be held in Lisbon, Portugal as part of IberSpeech 2016 (November 23-25, 2016). Representatives will be expected to give a presentation of their systems and to participate in discussions on the current state of the technology and future plans. The workshop will be open to participants in the ALBAYZIN 2016 Search on Speech Evaluation and to researchers registered to IberSpeech 2016.
- 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 IberSpeech 2016 website and emailed to the registered participants.

7 Acknowledgements

This work has been partly supported by projects CMC-V2 (TEC2012-37585-C02-01) and DSSL (TEC2015-68172-C2-1-P) from the Spanish Ministry of Economy and Competitiveness, and also by the Spanish Thematic Network on Speech Technology.

References

1. 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)

2. NIST: The spoken term detection (STD) 2006 evaluation plan. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 10 edn. (September 2006), http://www.nist.gov/speech/tests/std
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