The ALBAYZIN 2018 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 2018 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 2018 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 2018* to be held in Barcelona, Spain from 21 to 23 November 2018.[4]

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
[4] http://iberspeech2018.talp.cat

must generate a set of occurrences for each term detected in the audio files, along with their timestamps and scores as output. This is the same task as in NIST STD 2006 evaluation [1] and Open Keyword Search in 2013 [4], 2014 [5], 2015 [6], and 2016 [7].

– **Query-by-Example Spoken Term Detection (QbE STD)**, where the input to the system is an acoustic example per 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 scores as output, as in the STD task. This QbE STD is the same task as those proposed in MediaEval 2011, 2012, and 2013 [3].

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: an in-vocabulary (INV) set of terms/queries and an 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 4 contrastive systems for any task. Both development[5] 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.

---

[5] In case development data are provided for the corresponding database.

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 AL-BAYZIN Search on Speech evaluations, and comprises a set of talks extracted from the Spanish MAVIR workshops[6] held in 2006, 2007 and 2008 (Corpus MAVIR 2006, 2007 and 2008) corresponding to Spanish language, (2) COREMAH database[7][2], which contains recordings from native and non-native Spanish speakers with different Spanish levels, and was recorded in a university environment in 2014 and 2015, and (3) a database that comprises different Radio TeleVisión Española (RTVE) programs recorded from 2015 to 2018. For the MAVIR and the RTVE databases, three separate datasets (i.e., for training, development, and test) will be provided to participants. For COREMAH 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 the MAVIR and RTVE 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 [8]. The corresponding word transcription of the speech material will also be provided.

Regarding the RTVE database, about 492 hours of speech extracted from about 425 audio files 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[9] 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. The subtitles corresponding to these training data will also be provided, though these could not contain an accurate word transcription. The RTVE data are available

---

[6] http://www.mavir.net
[7] http://www.lllf.uam.es/coremah/
[8] http://sox.sourceforge.net/
[9] https://ffmpeg.org/

to the evaluation participants only and subject to the terms of a licence agreement with the RTVE. The license agreement can be downloaded from Cátedra RTVE-UZ web page (http://catedrartve.unizar.es/reto2018.html). Participants must sign the agreement and send a scanned copy attached to the email to *lleida@unizar.es* with the subject *RTVE2018 database license*. A copy signed by RTVE representative will be returned. Please read carefully the information provided on the Cátedra RTVE-UZ web page related with the use of the RTVE data after the evaluation campaign.

3.2 Development data

Development data belong to the same databases for which training data are provided (i.e., Spanish MAVIR workshop material and RTVE programs). 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.

Participants must submit an output result file for the development data corresponding to the MAVIR database and a different output result file for the development dataset named *dev2* corresponding to the RTVE 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 in the evaluation web page.

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 in the evaluation web page.

For the MAVIR database, the development list of terms consists of about 375 different terms (some of these are 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 are 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 development query list. The MAVIR development speech data amount to about 1 hour of speech material in total, extracted from 2 audio files. These MAVIR development data cannot be used for system training at all (in terms of acoustic models, language models, or in general, any kind of *training*).

For the RTVE database, two different datasets will be provided as development data: *dev1* and *dev2*. The dataset named *dev1* can be used as participants find more suitable (i.e., for training, development, etc.). The *dev1* RTVE development speech dataset amount to more than 60 hours of speech material in total, extracted from 51 audio files, for which human-revised word transcriptions will be provided. The dataset named *dev2* can only be used for development, and therefore cannot be used as training material at all. The *dev2* RTVE development speech dataset amount to about 15 hours of speech material in total, extracted from 12 audio files, for which human-revised word transcriptions will

be also provided. For this *dev2* dataset, the list of terms consists of about 400 different terms (some of these are INV terms and the rest are OOV) 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 (some of these are INV queries and the rest are OOV), with one example per query, will be extracted from the Spoken Term Detection list of terms to compose the Query-by-Example Spoken Term Detection development query list.

3.3 Test data

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

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 (some of these are 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 are 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 COREMAH data contain conversations between different topics such as rejection, compliment, and apology. Only the Spanish native speaker speeches are employed in the evaluation. The original speeches are videos in MPEG format. The audio of these videos was extracted and converted to PCM, 16khz, single channel, and 16 bits per sample using the *ffmpeg* tool[10]. The test speech data amount to about 45 minutes, extracted from 36 audio files. For the Spoken Term Detection task, only the list of terms used for evaluation will be provided. This list consists of about 175 different terms (some of these are INV terms and the rest are OOV) whose length ranges from 3 to 17 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 are 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 RTVE data consist of the RTVE program material explained before. The test speech data amount to about 10 hours in total, extracted from 24 audio files. For the Spoken Term Detection task, only the list of terms used for evaluation will be provided. This list consists of about 400 different terms (some of these are 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 are 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.

---

[10] https://ffmpeg.org/

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. The final ranking will be built from the average of the ATWVs obtained for the three databases on the test data (with equal weights for the three corpora). 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; 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 and RTVE 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.
- *RTVE2018/train/audio* - a folder with the RTVE training audio files.
- *RTVE2018/train/srt* - a folder with the subtitles of the RTVE training audio files. These may not contain accurate word transcriptions.
- 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 and RTVE 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 acoustic examples that serve as development queries for the Query-by-Example Spoken Term Detection task for the MAVIR database.
- *RTVE2018/dev1/audio* - a folder with the RTVE development *dev1* dataset audio files.

- *RTVE2018/dev1/trn* - a folder with the RTVE development *dev1* dataset human-revised word transcriptions.
- *RTVE2018/dev2/audio* - a folder with the RTVE development *dev2* dataset audio files.
- *RTVE2018/dev2/trn* - a folder with the RTVE development *dev2* dataset human-revised word transcriptions.
- *RTVE2018/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.
- *RTVE2018/dev2/queries* - a folder with the acoustic examples that serve as development queries for the Query-by-Example Spoken Term Detection task for the RTVE database.
- *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, COREMAH, and RTVE databases will consist of the following elements:

- *MAVIR/data* - a folder with the test audio files and two text files that contain the list of test terms/queries for the MAVIR database.
- *COREMAH/data* - a folder with the test audio files and two text files that contain the list of test terms/queries for the COREMAH database.
- *RTVE2018/test/audio* - a folder with the test audio files for the RTVE database.
- *RTVE2018/data* - a folder with two text files that contain the list of terms/queries to search for the RTVE database.
- *MAVIR/queries* - a folder with the acoustic examples that serve as test queries for the Query-by-Example Spoken Term Detection task for the MAVIR database.
- *COREMAH/queries* - a folder with the acoustic examples that serve as test queries for the Query-by-Example Spoken Term Detection task for the COREMAH database.
- *RTVE2018/queries* - a folder with the acoustic examples that serve as test queries for the Query-by-Example Spoken Term Detection task for the RTVE database.
- *scoring* - a folder with the scoring script, and the necessary input files for evaluating the systems for MAVIR, COREMAH, and RTVE 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 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 September 24th 2018, by contacting the organizing team at:

javiertejedornoguerales@gmail.com

with CC to ALBAYZIN 2018 Evaluations Organising Committee

albayzinevaluations@gmail.com

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 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, COREMAH, or RTVE 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 sent:

    HLPGA_STD_pri_DEV_MAVIR.xml
    HLPGA_STD_pri_DEV_RTVE.xml
    HLPGA_STD_pri_EVAL_MAVIR.xml
    HLPGA_STD_pri_EVAL_COREMAH.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_COREMAH.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_COREMAH.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_COREMAH.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_COREMAH.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 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 following the IberSpeech 2018 paper submission template. 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, 2018. Registration opens.
- June 30, 2018. Release of the training and development data.
- September 24, 2018. Release of the evaluation data. System submission opens. Registration deadline.
- October 21, 2018 (23:59, GMT +1). Deadline for the submission of results and system description paper.
- October 31, 2018. Results and ground-truth files are distributed to the participants.
- November 21-23, 2018. IberSpeech 2018, Barcelona, Spain: Evaluation results are presented and discussed.

6 Additional information for participants and summary of evaluation rules

– Interested groups must register for the evaluation before September 24th 2018, by contacting the organizing team at:

javiertejedornoguerales@gmail.com

with CC to ALBAYZIN 2018 Evaluations Organising Committee

albayzinevaluations@gmail.com

and providing the following information:

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

– Starting from June 30th 2018, and once registration data are validated, the training and development data will be released via web (only to registered participants).

– The test data will be released by September 24, 2018. Recognition results along with the system description paper must be submitted to the organizing team by the established deadline: October 21, 2018 (23:59 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 following the Iberspeech 2018 paper submission template.

– Registered groups commit themselves to use the provided data only for research purposes, distribution being allowed only with explicit permission of the ALBAYZIN 2018 Search on Speech Evaluation organizing team. Registered participants are allowed to use MAVIR and COREMAH 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",
"**COREMAH corpus**:
http://www.lllf.uam.es/coremah/",
and that they cite the Albayzin 2018 Search on Speech system description paper that will be included in the Iberspeech 2018 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 November 23rd 2018 (end of Iberspeech conference). Once this 3-year agreement finishes, participants can make a new request to continue using them. The request has to be made by mail with the following information: user, purpose of research, and period of use.

The address for the mail is as follows:

RTVE
Dir. FONDO DOCUMENTAL RTVE
Avda. Radiotelevisión, 4.
28223 Pozuelo de Alarcón
Madrid
España

Authors are also required to cite the Albayzin 2018 Search on Speech system description paper that will be included in the Iberspeech 2018 Proceedings in case of using the RTVE 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 Barcelona, Spain as part of IberSpeech 2018 (November 21-23, 2018). 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 2018 Search on Speech Evaluation and to researchers registered to IberSpeech 2018.

– 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 2018 website and emailed to the registered participants.

7   Acknowledgements

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. Matos, M.V.: Diseño y compilación de un corpus multimodal de análisis pragmático para la aplicación a la enseñanza del español. Ph.D. thesis, Universidad Autónoma de Madrid (2017)
3. 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)

4. 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
5. 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
6. 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
7. 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