Tejedor *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2017) 2017:22
DOI 10.1186/s13636-017-0119-z

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

ALBAYZIN 2016 spoken term detection evaluation: an international open competitive evaluation in Spanish

Javier Tejedor[1*], Doroteo T. Toledano[2], Paula Lopez-Otero[3], Laura Docio-Fernandez[4], Luis Serrano[5], Inma Hernaez[5], Alejandro Coucheiro-Limeres[6], Javier Ferreiros[6], Julia Olcoz[7] and Jorge Llombart[7]

Abstract

Within search-on-speech, Spoken Term Detection (STD) aims to retrieve data from a speech repository given a textual representation of a search term. This paper presents an international open evaluation for search-on-speech based on STD in Spanish and an analysis of the results. The evaluation has been designed carefully so that several analyses of the main results can be carried out. The evaluation consists in retrieving the speech files that contain the search terms, providing their start and end times, and a score value that reflects the confidence given to the detection. Two different Spanish speech databases have been employed in the evaluation: MAVIR database, which comprises a set of talks from workshops, and EPIC database, which comprises a set of European Parliament sessions in Spanish. We present the evaluation itself, both databases, the evaluation metric, the systems submitted to the evaluation, the results, and a detailed discussion. Five different research groups took part in the evaluation, and ten different systems were submitted in total. We compare the systems submitted to the evaluation and make a deep analysis based on some search term properties (term length, within-vocabulary/out-of-vocabulary terms, single-word/multi-word terms, and native (Spanish)/foreign terms).

Keywords: Search on speech, Spoken term detection, Spanish, International evaluation

1 Introduction

Search-on-speech aims to retrieve speech content from audio repositories that matches user queries. Due to the huge amount of information stored in audio and video repositories, the development of efficient methods for retrieving the stored information makes search-on-speech an important research area [1]. Within search-on-speech, there are several applications (that can be further divided into two different categories depending on the input/output of the system) shown in Table 1 for which significant research has been conducted. Within these applications, Spoken Term Detection (STD) is especially important, since this offers the possibility of retrieving any speech file that contains any term from its textual representation and hence is able to be used from devices with text input capabilities. Moreover, STD is also suitable for building open-vocabulary search-on-speech systems.

1.1 Spoken term detection overview

Spoken Term Detection has been receiving much interest for years from several companies/research institutes around the world (IBM [2–4], BBN [5], SRI and OGI [6–8], BUT [9–11], Microsoft [12], QUT [13, 14], JHU [15–17], Fraunhofer IAIS/NTNU/TUD [18], NTU [19, 20], IDIAP [21], etc). Spoken Term Detection systems are composed of two main stages: indexing by an Automatic Speech Recognition (ASR) subsystem, and then search by a detection subsystem, as depicted in Fig. 1. The ASR subsystem produces word/subword lattices from the input speech signal as an index. The detection subsystem integrates a term detector and a decision maker. The term detector searches for putative detections of the terms in the word/subword lattices, and the decision maker decides whether each detection is a hit or a false alarm (FA) based on certain confidence measures.

* Correspondence: javiertejedornogueralesgmail.com
[1]Escuela Politécnica Superior, Fundación Universitaria San Pablo CEU, Campus de Montepríncipe, Madrid, Spain
Full list of author information is available at the end of the article

Table 1 Search-on-speech applications

Type of result (output)	Type of query (input)	
	Text	Query
Document	SDR	QbE-SDR
Document+Position	STD/KWS	QbE-STD

'SDR' stands for Spoken Document Retrieval, 'STD' for Spoken Term Detection, 'KWS' for Keyword Spotting, and 'QbE' for Query-by-Example. STD is called when the audio is processed before knowing the terms to search, and KWS knows beforehand the terms to search

Word-based ASR has been widely used for building STD systems [7, 22–26] since this typically yields better performance than subword-based ASR [7, 8, 10, 22, 27–41] due to the lexical information the former employs. However, one of the main drawbacks of word-based ASR in STD is that this can only detect within-vocabulary, henceforth in-vocabulary (INV) terms. On the other hand, the subword-based approach has the unique advantage that it can detect terms that consist of words that are not in the recognizer's vocabulary (i.e., out-of-vocabulary (OOV) terms). In order to exploit the relative advantages of the word and subword-based approaches, it has been proposed to combine these two approaches [8, 11, 12, 22, 38, 42–49].

The availability of several ASR tools (e.g., Hidden Markov Model Toolkit (HTK) [50], Sphinx [51], Kaldi [52] etc.) facilitates the development of STD systems, since these save researchers constructing an ASR system from scratch. Among these, Kaldi is especially suitable for building STD systems, since this integrates an ASR subsystem, a term detector, and a decision maker [52–54]. The Kaldi STD system employs a word-based approach for term detection, and a method based on proxy-words (i.e., replacing each OOV word by the most similar in-vocabulary word) to detect OOV terms [55].

1.2 Motivation and organization of this paper

In general, the systems developed for STD research are difficult to compare since just a few of them are typically evaluated under a common framework (i.e., the speech database and the search terms differ from one to another). International evaluations, however, provide a framework that can be effectively employed to evaluate the progress of the technology.

Specifically, ALBAYZIN evaluation campaigns comprise an international open set of evaluations supported by the Spanish Thematic Network on Speech Technologies (RTTH[1]) and the ISCA Special Interest Group on Iberian Languages (SIG-IL[2]), which have been held every 2 years since 2006. These evaluation campaigns provide an objective mechanism to compare different systems and are a powerful way to promote research on different speech technologies [56–63].

Spanish is a major language in the world and significant research has been conducted on it for ASR, KeyWord Spotting (KWS), and STD tasks [64–70]. The increasing interest in search-on-speech around the world and the lack of search-on-speech evaluations dealing with Spanish language encouraged us to organize a series of STD evaluations starting in 2012 and held biannually until 2016 aiming to evaluate the progress in this technology for Spanish. Each evaluation has focused on improving its strategy by incorporating new challenges. This third ALBAYZIN STD evaluation was specifically designed to evaluate the performance of STD systems that address several challenging conditions (in-vocabulary vs. out-of-vocabulary terms, native (Spanish), henceforth in-language (INL) vs. out-of-language (OOL) terms, single-word vs. multiple-word terms, and single-domain vs. multiple-domain databases).

This evaluation also incorporated stricter rules regarding the evaluation terms. In addition, all the terms and the database employed in the STD evaluation held in 2014 were kept, enabling a comparison between the systems submitted to both evaluations on the common set of terms.

The rest of the paper is organized as follows: next section presents the STD evaluation, the metric used, the databases released for experimentation, a comparison with previous evaluations, and the participants involved in the evaluation. Then, in Section Systems, we present the

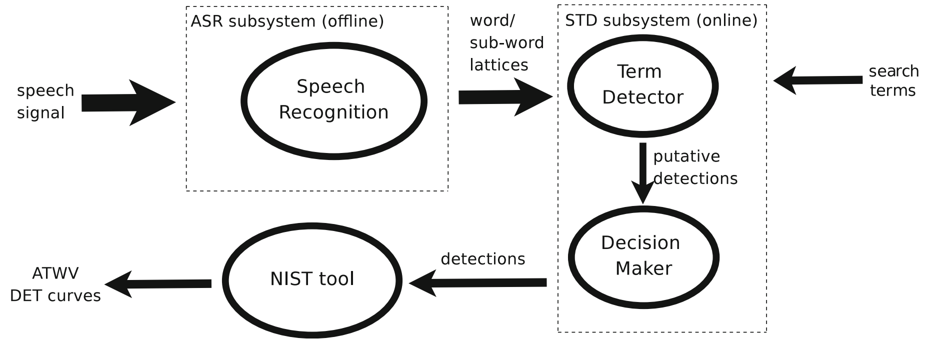

Fig. 1 STD system architecture

different systems submitted for the evaluation. Results along with discussion are presented in Section Results and discussion and the paper is concluded in the last section.

2 Spoken term detection evaluation

2.1 STD evaluation overview

This evaluation involves searching a list of terms within speech content. In other words, the STD evaluation focuses on retrieving the appropriate audio files, with the occurrences and timestamps, which contain any of those terms.

The evaluation consists of searching a development term list within development speech data, and searching two different test term lists within two different sets of test speech data (MAVIR and EPIC databases as can be seen next). The evaluation result ranking is based on the system performance when searching the test terms within the test speech data corresponding to the MAVIR database. Training data can only be used for participants for system training. Development data can only be used for participants for system tuning. Additionally, any other data could also be used for participants for both stages (training and development).

This evaluation defined two different sets of terms for each database: the in-vocabulary term set and the out-of-vocabulary term set. The OOV term set was defined to simulate the out-of vocabulary words of a Large Vocabulary Continuous Speech Recognition (LVCSR) system. In case participants employ an LVCSR system for processing the audio, these OOV terms must be previously removed from the system dictionary and hence, other methods have to be used for searching OOV terms. On the other hand, the INV terms could appear in the LVCSR system dictionary in case participants consider it.

Participants could submit a primary system and up to two contrastive systems. No manual intervention was allowed for each developed system to generate the final output file and hence, all the developed systems had to be fully automatic. Listening to the test data, or any other human interaction with the test data was forbidden before all the evaluation results in terms of the performance of the systems in test data (i.e., evaluation result ranking) had been sent back to the participants. The standard eXtensible Markup Language (XML)-based format corresponding to the National Institute of Standards and Technology (NIST) evaluation tool [5] was used for building the system output file. Ground-truth labels corresponding to the test data were given to participants once the Organizers sent them back the evaluation results.

2.2 Evaluation metric

In STD, a hypothesized occurrence is called a detection; if the detection corresponds to an actual occurrence, it is called a hit, otherwise it is a false alarm. If an actual occurrence is not detected, this is called a miss. The Actual Term Weighted Value (ATWV) proposed by NIST [5] has been used as the main metric for the evaluation. This metric integrates the hit rate and false alarm rate of each term into a single metric and then averages over all the terms:

$$\mathrm{ATWV} = \frac{1}{|\Delta|}\sum_{K\in\Delta}\frac{N^K_{\mathrm{hit}}}{N^K_{\mathrm{true}}} - \beta\frac{N^K_{\mathrm{FA}}}{T - N^K_{\mathrm{true}}}, \quad (1)$$

where Δ denotes the set of terms and $|\Delta|$ is the number of terms in this set. N^K_{hit} and N^K_{FA} represent the numbers of hits and false alarms of term K, respectively, and N^K_{true} is the number of actual occurrences of K in the audio. T denotes the audio length in seconds (i.e., the number of seconds of the corresponding speech files where the terms are searched) and β is a weight factor set to 999.9, as in the ATWV proposed by NIST [5]. This weight factor causes an emphasis placed on recall compared to precision in the ratio 10:1.

ATWV represents the term weighted value (TWV) for the threshold set by the STD system (usually tuned on development data). An additional metric, called maximum term weighted value (MTWV) [5] can also be used to evaluate the performance of an STD system. This MTWV is the maximum TWV achieved by the STD system for all possible thresholds, and hence does not depend on the tuned threshold. Therefore, this MTWV represents an upper bound of the performance obtained by the STD system. Results based on this metric are also presented to evaluate the system performance with respect to threshold selection.

In addition to ATWV and MTWV, NIST also proposed a Detection Error Tradeoff (DET) curve [71] to evaluate the performance of an STD system working at various miss/FA ratios. Although DET curves were not used for the evaluation itself, they are also presented in this paper for system comparison.

The NIST STD evaluation tool [72] was employed to compute MTWV, ATWV, and DET curves.

2.3 Databases

Two different databases that comprise different acoustic conditions and domains have been employed for the evaluation. For comparison purposes, the same MAVIR database employed in the ALBAYZIN STD evaluation held in 2014 has been used. The second database is the EPIC database distributed by ELRA[3]. For MAVIR database, three separate datasets (i.e., for training, development, and test) were provided to participants. For EPIC database, only test data were provided. This allowed measuring the generalization capability of the systems in an unseen domain.

MAVIR database consists of a set of Spanish talks extracted from the MAVIR workshops[4] held in 2006, 2007, and 2008 (corpus MAVIR 2006, 2007, and 2008) that contain speakers from Spain and Latin America.

The MAVIR Spanish data consist of spontaneous speech files, each containing different speakers, which amount to about 7 h of speech and are further divided for the purpose of this evaluation into training, development, and test sets. The data were also manually annotated in an orthographic form, but timestamps were only set for phrase boundaries. To prepare the data for the evaluation, organizers manually added the timestamps for the roughly 3000 occurrences of the spoken terms used in the development and test evaluation sets. The training data were made available to the participants including the orthographic transcription and the timestamps for phrase boundaries[5].

The speech data were originally recorded in several audio formats (Pulse Code Modulation (PCM) mono and stereo, Moving Picture Experts Group (MPEG)-1 Audio Layer 3 (MP3), 22.05 KHz., 48 KHz., etc.). All data were converted to PCM, 16 KHz., single channel, 16 bits per sample using SoX tool[6]. Recordings were made with the same equipment, a Digital TASCAM DAT model DA-P1, except for one recording. Different microphones were used for the different recordings. They mainly consisted of tabletop or floor standing microphones, but in one case a lavalier microphone was used. The distance from the mouth of the speaker to the microphone varies and was not particularly controlled, but in most cases the distance was smaller than 50 cm. All the recordings contain spontaneous speech of MAVIR workshops in a real setting. Thus, the recordings were made in large conference rooms with capacity for over a hundred people and a large amount of people in the conference room. This poses additional challenges including background noise (particularly babble noise) and reverberation. The realistic settings and the different nature of the spontaneous speech in this database make it appealing and challenging enough for the evaluation. Table 2 includes some database features such as the division in training, development, and test data of the speech files, the number of word occurrences, duration, and p.563 Mean Opinion Score (MOS) [73] as a way to get an idea of the quality of each speech file in the MAVIR Spanish database. The p.563 MOS algorithm is intended to be employed for evaluating the quality of the human voice. In addition, no reference signal is needed to compute the p.563 MOS value. MOS values are in the range of 1–5, being 1 the worst quality and 5 the best quality. More information about the p.563 algorithm can be found in [73].

EPIC database comprises speech data from the European Parliament Interpretation Corpus recorded in 2004 in English, Spanish, and Italian, along with their corresponding simultaneous interpretations to the other languages. Only the Spanish original speeches, which amount to more than 1:5 h of speech, were used for the evaluation. Table 3 includes some features of EPIC database as those shown in Table 2.

2.3.1 Term list selection

All the terms selected for both development and test sets aimed to build a realistic scenario for STD, by including high occurrence terms, low occurrence terms, foreign terms, single-word and multi-word terms, in-vocabulary and out-of-vocabulary terms, and different length terms. A term may not have any occurrence or appear one or more times in the speech data. Table 4 includes some features of the development and test term

Table 2 MAVIR database characteristics

File ID	Data	#word occ.	dur. (min)	#spk.	p.563 Ave. MOS
Mavir-02	train	13432	74.51	7 (7 ma.)	2.69
Mavir-03	dev	6681	38.18	2 (1 ma. 1 fe.)	2.83
Mavir-06	train	4332	29.15	3 (2 ma. 1 fe.)	2.89
Mavir-07	dev	3831	21.78	2 (2 ma.)	3.26
Mavir-08	train	3356	18.90	1 (1 ma.)	3.13
Mavir-09	train	11179	70.05	1 (1 ma.)	2.39
Mavir-12	train	11168	67.66	1 (1 ma.)	2.32
Mavir-04	test	9310	57.36	4 (3 ma. 1 fe.)	2.85
Mavir-11	test	3130	20.33	1 (1 ma.)	2.46
Mavir-13	test	7837	43.61	1 (1 ma.)	2.48
ALL	train	43467	260.27	13 (12 ma. 1 fe.)	-
ALL	dev	10512	59.96	4 (3 ma. 1 fe.)	-
ALL	test	20277	121.3	6 (5 ma. 1 fe.)	-

'train' stands for training, 'dev' for development, 'occ.' stands for occurrences, 'dur.' stands for duration, 'min' stands for minutes, 'spk.' stands for speakers, 'ma.' stands for male, 'fe.' stands for female, and 'Ave.' stands for average

Table 3 EPIC database characteristics

File ID	#word occ.	dur. (min)	#spk.	p.563 Ave. MOS
10-02-04-m-058-org-es	280	2.47	1 fe.	3.71
10-02-04-m-074-org-es	3189	25.2	1 ma	2.79
11-02-04-m-017-org-es	532	3.47	1 fe.	3.70
11-02-04-m-022-org-es	896	5.08	1 ma.	2.76
11-02-04-m-032-org-es	726	3.37	1 ma.	3.12
11-02-04-m-035-org-es	535	3.12	1 ma.	3.44
11-02-04-m-041-org-es	92	0.78	1 ma.	3.00
11-02-04-m-054-org-es	199	1.70	1 ma.	3.12
12-02-04-m-010-org-es	344	2.38	1 ma.	3.18
12-02-04-m-028-org-es	78	0.45	1 ma.	1.66
12-02-04-m-038-org-es	285	2.17	1 ma.	3.31
25-02-04-p-024-org-es	1205	8.50	1 fe.	3.92
25-02-04-p-027-org-es	353	2.23	1 ma.	3.67
25-02-04-p-030-org-es	523	3.18	1 fe.	3.79
25-02-04-p-034-org-es	353	2.23	1 fe.	3.78
25-02-04-p-037-org-es	492	2.93	1 fe.	3.67
25-02-04-p-043-org-es	1705	12.27	1 ma.	3.32
25-02-04-p-047-org-es	922	5.82	1 ma.	3.39
25-02-04-p-072-org-es	278	1.90	1 fe.	4.27
25-02-04-p-081-org-es	1270	8.07	1 ma.	3.20
25-02-04-p-096-org-es	211	1.27	1 ma.	3.41
ALL	14468	98.58	21 (14 ma. 7 fe.)	-

'occ.' stands for occurrences, 'dur.' stands for duration, 'min' stands for minutes, 'spk.' stands for speakers, 'ma.' stands for male, 'fe.' stands for female, and 'Ave.' stands for average

lists such as the number of in-language and foreign terms, the number of single-word and multi-word terms, and the number of in-vocabulary and out-of-vocabulary terms, along with the number of occurrences of each set in the corresponding speech data. It must be noted that a multi-word term is considered OOV in case any of the words that form the term is OOV. It can be seen that the test EPIC term list only contains easy terms (i.e., no out-of-language and multi-word terms are included), since the main purpose was to evaluate the systems submitted to the evaluation in a different domain.

Table 4 Development and test term list characteristics for MAVIR and EPIC databases

Term list	dev	test-MAVIR	test-EPIC
#IN-LANG terms (occ.)	354 (959)	208 (2071)	183 (1912)
#OUT-LANG terms (occ.)	20 (55)	15 (50)	0 (0)
#SINGLE terms (occ.)	340 (984)	198 (2093)	183 (1912)
#MULTI terms (occ.)	34 (30)	25 (28)	0 (0)
#INV terms (occ.)	292 (668)	192 (1749)	150 (1562)
#OOV terms (occ.)	82 (346)	31 (372)	33 (350)

'dev' stands for development, 'IN-LANG' refers to in-language terms, 'OUT-LANG' to foreign terms, 'SINGLE' to single-word terms, 'MULTI' to multi-word terms, 'INV' to in-vocabulary terms, 'OOV' to out-of-vocabulary terms, and 'occ.' stands for occurrences. The term length of the development term list varies between 5 and 27 graphemes. The term length of the MAVIR test term list varies between 4 and 28 graphemes. The term length of the EPIC test term list varies between 6 and 16 graphemes

2.4 Comparison to other STD international evaluations

In 2006, the National Institute of Standards and Technology of the United States of America (USA) organized the first NIST STD evaluation [5], in which English, Mandarin Chinese, and Modern Standard and Levantine Arabic languages were included. The speech used in this evaluation included conversational telephone speech (CTS), broadcast news (BNews) speech, and speech recorded in roundtable meeting rooms (RTMeet) with distantly placed microphones (this last type was used for English only). The NIST STD 2006 evaluation results are publicly available[7], and are very interesting to analyze the influence of the language and the type of speech on STD results. Table 5 presents the best results obtained by the participants for each condition. With respect to the type of speech, it is clear from Table 5 that results using microphone speech, particularly distant microphones in less controlled settings than in audiovisual

Table 5 Best performance (in terms of Actual Term Weighted Value, ATWV) obtained by the different participants of the NIST STD 2006 evaluation in the different conditions: 'CTS' stands for Conversational Telephone Speech, 'BNews' for Broadcast News, and 'RTMeet' for speech recorded in roundtable meeting rooms

Language	CTS	BNews	RTMeet
English	0.8335	0.8485	0.2553
Arabic	0.3467	-0.0924	N/A
Mandarin	0.3809	N/A	N/A

studios (such as in broadcast news) or close-talking conversational telephone data, are definitely much more limited. With respect to the language, English is the language with more resources and for which more research has been done so far. When applying similar technology to languages for which less specific research has been conducted, performance decreases are observed.

A significant amount of research has been carried out in the framework of the IARPA BABEL program and NIST Open Keyword Search (OpenKWS) evaluation series [74–84]. BABEL program was born in 2011 aiming to develop fully automatic and noise-robust speech recognition systems in limited time (e.g., 1 week) with limited amount of transcribed training data. This program supports research in low-resource languages such as Cantonese, Pashto, Tagalog, Turkish, Vietnamese, etc. Since 2013, NIST has been organizing an annual open STD evaluation called OpenKWS, which is included within the BABEL program, but open to other research groups besides BABEL participants [85–88]. This evaluation is very similar to the former NIST STD 2006 evaluation and aimed to build STD systems in a limited time for low-resource languages. This includes CTS and microphone speech data on a surprise language that was announced only a few (4 or less) weeks before the evaluation. The main results of these OpenKWS evaluations are shown in Table 6.

In our evaluation, the audio contains microphone recordings of real talks in real workshops, in large conference rooms in public. Microphones, conference rooms, and even recording conditions change from one recording to another. Microphones are not close-talking microphones but mainly tabletop and ground standing microphones. In addition, our evaluation explicitly defines different in-vocabulary and out-of-vocabulary set of terms. These differences in the evaluation conditions make our evaluation pose different challenges, and make it difficult to compare the results obtained in our evaluation to those of the previous NIST STD evaluations.

Table 6 Best performance (in terms of Actual Term Weighted Value, ATWV) obtained by the different participants in the OpenKWS evaluations held in 2013, 2014, and 2015 under the full language pack condition

Evaluation	ATWV	Language
OpenKWS 2013	0.6248	Vietnamese
OpenKWS 2014	0.5802	Tamil
OpenKWS 2015	0.6548	Swahili

STD evaluations have also been held in the framework of the NTCIR conferences [89–92]. Data used in these evaluations contained spontaneous speech in Japanese provided by the National Institute for Japanese language and spontaneous speech recorded during seven editions of the Spoken Document Processing Workshop. These evaluations also provide the manual transcription of the speech data and the output of an LVCSR system to the participants. Table 7 presents the best result obtained in each individual evaluation, where the F-measure was used as the evaluation metric. Although our evaluation could be similar in terms of speech nature to these NTCIR STD evaluations (speech recorded in real workshops), our evaluation makes use of other language, employs a larger list of terms along with two different databases, and defines disjoint development and test term lists to measure the generalization capability of the systems. Besides, the evaluation metric used by these evaluations is different, which makes comparison very difficult.

2.5 Participants

Ten different systems were submitted from five different research groups to the ALBAYZIN Spoken Term Detection 2016 evaluation, as listed in Table 8. About 3 months were given to the participants for system development and hence, the STD evaluation focuses on building STD systems in limited time. The training, development, and test data were released to the participants in different periods. Training and development data were released by the end of June 2016. The test data were released at the beginning of September 2016. The final system submission was due by mid-October 2016. Final results were discussed at IberSPEECH 2016 conference by the end of November 2016.

3 Systems

In this section, the systems submitted to the evaluation are described. All the systems integrate two subsystems: an ASR subsystem based on the Kaldi open-source toolkit

Table 7 Best performance (in terms of F-measure) obtained by the different participants in the NTCIR STD evaluations

Evaluation	F-measure
NTCIR STD-09	0.3660
NTCIR STD-10	0.7944
NTCIR STD-11	0.6140
NTCIR STD-12	0.7188

Table 8 Participants in the ALBAYZIN Spoken Term Detection 2016 evaluation along with the systems submitted

Team ID	Research Institution	Systems
GTM-UVigo	AtlantTIC Research Center Universidad de Vigo, Spain	Combined Kaldi Proxy Kaldi
AHOLAB	Universidad del País Vasco, Spain	Syn-Syll Comb
ATVS-FOCUS	Universidad Autónoma de Madrid - FOCUS, Spain	DD Kaldi DI Kaldi
GTH-UPM	Universidad Politécnica de Madrid, Spain	Kaldi matcher
ViVOLAB	Universidad de Zaragoza, Spain	Phone STD Mismatch Phone STD CM Phone STD CM Phone STD+TST

[52], and an STD subsystem that comprises a term detector and a decision maker.

3.1 Combined Kaldi-based STD system (combined Kaldi)

This system consists of the combination of a word-based STD system to detect INV words, and a phone-based STD system to detect OOV words, as depicted in Fig. 2. Both systems are described next.

3.1.1 Word-based STD system

The ASR subsystem is based on the Kaldi open-source toolkit [52] and employs deep neural network (DNN)-based acoustic models. Specifically, a DNN-based context-dependent speech recognizer was trained following the DNN training approach presented in [93]. 40-dimensional Mel-frequency cepstral coefficients (MFCCs) augmented with three pitch and voicing related features [94], and appended with their delta and acceleration coefficients were first extracted for each speech frame. The DNN has six hidden layers with 2048 neurons each. Each speech frame is spliced across ±5 frames to produce 1419-dimensional vectors which are the input to the first layer, and the output layer is a soft-max layer representing the log-posteriors of the context-dependent Hidden Markov Model (HMM) states. The Kaldi LVCSR decoder generates word lattices [95] using these DNN-based acoustic models.

The data used to train the acoustic models of this Kaldi-based LVCSR system were extracted from the Spanish material used in the 2006 TC-STAR automatic speech recognition evaluation campaign[8] and the Galician broadcast news database Transcrigal [96]. It must be noted that all the non-speech parts as well as the speech parts corresponding to transcriptions with pronunciation errors, incomplete sentences, and short speech utterances were discarded, so in the end the acoustic training material consisted of approximately 104.5 h.

The language model (LM) employed in the LVCSR system was constructed using a text database of 160 million of word occurrences composed of material from several sources (transcriptions of European and Spanish Parliaments from the TC-STAR database, subtitles, books, newspapers, on-line courses, and the transcriptions of the MAVIR sessions included in the development set provided by the evaluation organizers[9] [97]). Specifically, the LM was obtained by static interpolation of trigram-based language models which were trained using these different text databases. All LMs were built using the Kneser-Ney discounting strategy using the SRILM toolkit [98], and the final interpolated LM was obtained using the SRILM static n-gram interpolation functionality. The LM vocabulary size was limited to the most frequent 60,000 words and, for each evaluation data set, the OOV terms were removed from the language model.

The STD subsystem integrates the Kaldi term detector [52–54] which searches for the input terms within the word lattices obtained in the previous step. To do so, these lattices are processed using the lattice indexing technique described in [99] so that the lattices of all the utterances in the search collection are converted from individual weighted finite state transducers to a single generalized factor transducer structure in which the start-time, end-time, and lattice posterior probability of each word token are stored as three-dimensional costs. This factor transducer is actually an inverted index of all word sequences seen in the lattices. Thus, given a list of terms, a simple finite state machine is created such that it accepts each term and composes it with the factor transducer to obtain all occurrences of the terms in the search collection. The Kaldi decision maker conducts a YES/NO decision for each detection based on the term specific threshold (TST) approach presented in [23]. To

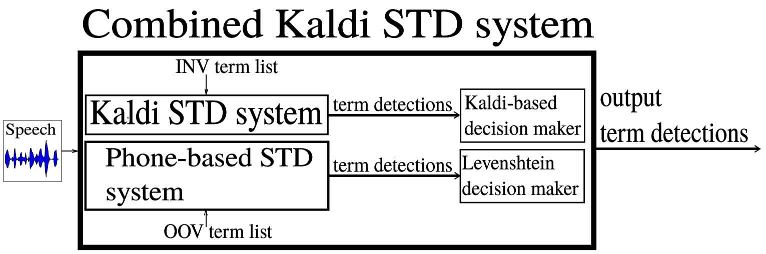

Fig. 2 Combined Kaldi-based STD system architecture

do so, the score for each detection is computed as follows:

$$p > \frac{N_{\text{true}}}{\frac{T}{\beta} + \frac{\beta - 1}{\beta} N_{\text{true}}}, \tag{2}$$

where p is the confidence score of the detection, N_{true} is the sum of the confidence score of all the detections of the given term, β is set to 999.9, and T is the length of the audio in seconds.

3.1.2 Phone-based STD system

OOV terms were handled with a phone-based STD system strategy. A phoneme sequence was first obtained from the 1-best word path obtained with the word-based Kaldi LVCSR system presented before. Next, a reduction of the phoneme set was performed in order to combine phonemes with high confusion; specifically, semivowels /j/ and /w/ were represented as vowels /i/ and /u/, respectively, and palatal n /ŋ/ was represented as /n/. Then, the tre-agrep tool is employed to compute partial matches so that the Levenshtein distance between each recognized phoneme sequence and the phoneme sequence corresponding to each term can be computed. An analysis of the proposed strategy suggested that those matches whose Levenshtein distance was equal to 0 were, in general, correct matches. Matches with Levenshtein distance equal to 1 were more prone to be false alarms, although many matches were found as well; since no specific criteria to assign a score was implemented, only those matches with Levenshtein distance equal to 0 were kept, and they were assigned the maximum score (1). The OOV term detections found using this phone-based STD system were directly merged with the INV detections obtained using the word-based STD system.

3.2 Proxy Kaldi STD system (proxy Kaldi)

This system, whose architecture is presented in Fig. 3, is the same as the Combined Kaldi system, except that the OOV terms were searched with the proxy words strategy of the Kaldi toolkit [55]. This strategy consists of substituting each OOV word of the search term with acoustically similar INV proxy words, getting rid of the need of a subword-based system for OOV term search.

3.3 Synthesis and syllabic decomposition-based combined STD system (Syn-Syll comb)

This system, whose architecture is presented in Fig. 4, integrates different strategies in the STD component.

The ASR component is based in Kaldi and follows the s5 recipe for the Wall Street journal database. The acoustic features used are 13-dimensional MFCCs with cepstral mean and variance normalization (CMVN). These features are first used to train context-independent Gaussian Mixture Model (GMM)-HMMs. Next, context-dependent GMM-HMMs have been trained. After that, a new set of features is derived by processing the initial 13-dimensional MFCCs. Specifically, these 13-dimensional features are spliced across 4 frames to produce 117-dimensional vectors. Next, Linear Discriminant Analysis (LDA) is used to reduce the dimensionality to 40, and a maximum likelihood linear transform (MLLT) is applied to match the diagonal assumption in GMM. The context-dependent GMM-HMM states are used as classes for the LDA estimation. Next, feature-based maximum likelihood linear regression (fMLLR) and speaker adaptive training (SAT) techniques are applied to improve model robustness. With this MFCC (spliced) + LDA + MLLT + fMLLR + SAT 40-dimensional features, DNN-based acoustic models were trained. The DNN is trained with four hidden layers, and follows the training approach (layer-wise back-propagation) described in [100].

The DNN-based acoustic models were trained from a set of 47 parliamentary sessions that took place in the Basque parliament[10], of which only the Spanish part was employed. As a result, more than 124 h of speech in Spanish, uttered by 45 male and 39 female speakers, were employed. In addition, the training data provided by the organizers were also employed for acoustic model training. These data consist of 4 h of spontaneous speech.

A dictionary composed of more than 37,000 words has been constructed for the ASR component. In addition, trigram and uniform unigram LMs have been built. The trigram LM has been trained from a set of 54 million of words (after removing the OOV words) corresponding to the European Parliament Proceedings Parallel Corpus 1996–2011 [101]. On the other hand, the uniform unigram LM has been trained using the training input dictionary created with the material provided by the organizers. The SRILM toolkit [98] has been used to train both LMs. Two different LVCSR processes were conducted: The first one employs the trigram LM, and the second one employs the uniform unigram LM. Both produce word lattices that comprise the input to the STD subsystem.

For the STD subsystem, the INV terms were searched from the word lattices obtained in the ASR component using the Kaldi term detector [52–54]. A two-pass strategy was employed. In the first-pass, the Kaldi term detector receives as input the lattices obtained in the trigram LM-based LVCSR decoding and produces putative detections. These putative detections are taken by the Kaldi decision maker and sorted according to their posterior probability to assign a preliminary YES/NO decision to each detection. This decision is next modified as follows: if the number of occurrences of a certain term is above a predefined threshold t, all the occurrences for this term that present a probability higher than a certain score s are assigned YES

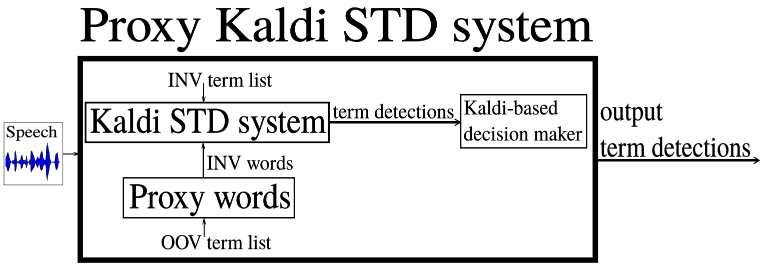

Fig. 3 Proxy Kaldi STD system architecture

decision. For the terms that have not been detected in the first pass, a second pass is conducted. In this second pass, the lattices given to the Kaldi term detector were obtained from the uniform unigram LM, and the YES/NO decision is given by the Kaldi decision maker from the posterior probability assigned to each detection.

The purpose of this second pass is to minimize the effect of the trigram LM and favor the acoustic models. The occurrences of the terms detected in this second pass are added to those obtained in the first pass.

The OOV terms were handled using the proxy words strategy of the Kaldi toolkit [55]. As for INV terms, a two-pass strategy has been applied: In the first pass, the OOV terms were synthesized and recognized to create proxy word Finite State Transducers (FSTs) that are needed in the Kaldi term detector to produce a first set of OOV term detections. In the second pass, all the OOV terms with no occurrences in the first pass were searched by means of a syllabic decomposition to produce a second set of OOV term detections. All the detections corresponding to both OOV term sets were merged to the INV term detections to produce the final list of detections. Next, these two passes are explained.

3.3.1 Text to speech synthesis-based first pass

All the OOV terms are synthesized using the Aholab Text-to-Speech synthesizer [102]. The generated synthetic signals are then given to the Kaldi-based LVCSR system, where the uniform unigram LM was employed to obtain the word-based lattices. From the word lattices, the best hypothesis is chosen as the term to search, i.e., as proxy term. The aimed goal is to get the most acoustically similar INV term for each OOV term. Next, each FST corresponding to each OOV term is built. The resulting FSTs are given to the Kaldi term detector, which uses the lattices obtained from the uniform unigram LM to hypothesize detections. Next, the Kaldi decision maker assigns the YES/NO decision to each detection. The possibility of using more than one hypothesis was discarded because the results showed a high number of false alarms.

3.3.2 Syllabic decomposition-based second pass

First, the audio is recognized using the Kaldi-based LVCSR system with the uniform unigram LM, and the 1-best path is stored. Then, the corresponding 1-best path transcription that contains the start-time and end-time of each word recognized is obtained. The next step is to decompose the

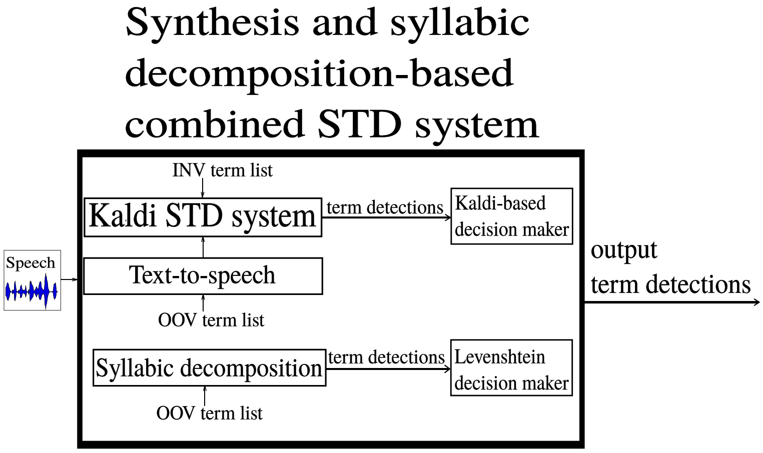

Fig. 4 Synthesis and Syllabic decomposition-based combined STD system architecture

words of the transcription in syllables. The OOV terms are decomposed in syllables too, and a measure of the difference between the syllables of each term and the syllables of the whole transcription is calculated. A window with the length of the syllables of the OOV term slides through the transcription text and the difference is calculated based on the phonetic transcription of the syllables. Every place where this difference falls below a predefined threshold is taken as an occurrence of the search term. The final score obtained for each search window is calculated as 1 - dL where dL represents the Levenshtein distance. In our case, the penalty assigned to an insertion/deletion of a phone of the syllable is set to 0.5, and the penalty for replacing one phone by another can vary from 0 to 1 according to the previously estimated acoustic distance between the two phones. This penalty was estimated from the data used for acoustic model training as follows:

- Training data are given to the HTK tool [50] to train HMM monophone models. Each HMM consist of 39-dimensional, three emitting-states for each phone (24 in this case).
- The mean vector of the central state of each phone model is selected, and the Euclidean distance between all the mean vectors is calculated to have an estimation of the acoustic distance.
- The resulting distances are normalized to the maximum distance. The distance computed between a phone and itself is set to 0.

Once the distance is calculated, a decision for each detection must be taken (i.e., some distance threshold must be set). The threshold value to consider a detection as a true occurrence was chosen empirically based on the mean score and the standard deviation of each detection evaluated over the whole transcription. This threshold was selected from the development data provided by the organizers to be quite conservative (i.e., to minimize the insertion of false alarms).

3.4 Domain dependent Kaldi-based STD system (DD Kaldi)

This system is the standard Kaldi STD system with its decision maker, and its structure is the same as that in Fig. 3.

For the ASR subsystem, first an energy-based Voice Activity Detector (VAD) implemented in SoX removed non-speech segments. For word-based speech recognition, the system used the s5 recipe of the Kaldi toolkit [52] for the Fisher Callhome Spanish database. To do so, 13-dimensional MFCCs with cepstral mean and variance normalization were used as acoustic features. The normalized MFCC features then pass a splicer which augments each frame by its left and right 4 neighboring frames. LDA was applied to reduce the feature dimension to 40, and next, MLLT was also applied. fMLLR and SAT were also applied to improve model robustness. A subspace Gaussian Mixture Model (SGMM) was then trained, and a discriminative training approach based on boosted maximum mutual information (bMMI) is used to produce better acoustic models.

The SGMM models were trained with the training part of the Fisher Spanish corpus and the Callhome Spanish corpus (about 268 h of conversational telephone speech in total), and the training data provided by the organizers (about 4 h of speech). For English words that appear in the Fisher Spanish corpus, a letter-to-sound module has been used to build the phone transcription of these words using Spanish phones. To do so, the Carnegie Mellon University (CMU) Dictionary was employed to obtain an English phoneme transcription and define translation rules from English to Spanish phonemes to build the phoneme transcription of the English words using Spanish phonemes. Interjections and acronyms were transcribed manually. In the end, a dictionary of about 36,000 terms, fully transcribed with a set of 24 Spanish phonemes with stress marked as different phonemes, was built. Besides the phoneme models, models for laughter and noise were included.

A trigram LM was employed for ASR decoding. This LM was trained from the SRILM toolkit [98] using the word transcription of the data employed for acoustic model training.

As in the Combined Kaldi system explained before, the Kaldi term detector [52–54] was employed to hypothesize detections from the word lattices obtained in the ASR component, and the Kaldi decision maker ascertains reliable detections.

OOV term detection has been carried out using the *proxy words* mechanism of the Kaldi toolkit [55], as in the Proxy Kaldi system explained before.

3.5 Domain independent Kaldi-based STD system (DI Kaldi)

This system is the same as the previous one (DD Kaldi) except that the training data provided by the organizers were not employed for acoustic model and language model training.

3.6 Kaldi-word + phone matcher decoding STD system (Kaldi matcher)

This system consists of two speech recognition processes: a word-based speech recognition, and a phone-based speech recognition, and different term detection modules, as shown in Fig. 5.

The ASR component integrates word and phone-based speech recognizers built with the Kaldi toolkit [52]. First, feature vectors consisting of 13 MFCCs and their first and second order time derivatives are used to train tied-pdf cross-word triphone context, three hidden state phone models. This results in GMM-HMMs with 200 k Gaussians

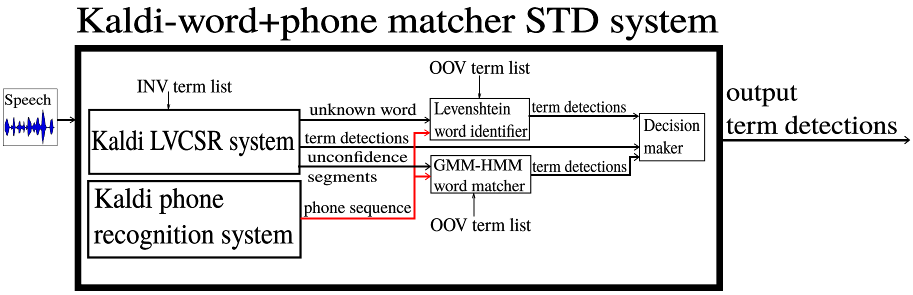

Fig. 5 Kaldi - word + phone matcher decoding STD system architecture

and 4294 senones. Then, feature-level transformations such as LDA, MLLT, and fMLLR were applied. This results in 40-dimensional acoustic vectors per frame of 10 ms (Fig. 6). The GMM-HMMs served as alignment sources for training a DNN-HMM model, with 4 hidden layers in a 2-norm maxout network [100] with 3000 nodes per hidden layer with groups of 10. The number of spliced frames was 9. The network was trained along 20 epochs.

DNN-HMM ASR acoustic models were trained from the training data provided by the organizers, and the Spanish partition of the *European Parliament Plenary Sessions* (EPPS) and the *CONG-PARL* databases included in the TC-STAR database [103]. As for the LMs, different LMs were used for the word and phone-based speech recognizers, and for the different evaluation data sets. The word-based LM consists of a trigram LM trained from different text sources, which vary for MAVIR or EPIC data processing: for MAVIR development data processing, the data employed for LM training comprise the training data provided by the organizers and web texts (which amount to about 400 k word occurrences) related to similar topics to those of the MAVIR training data; for MAVIR test data processing, the trigram LM was enhanced with the MAVIR development data transcription; for EPIC data processing (test), the LM training data comprise the transcription data corresponding to 100 h of speech of the EPPS and CONG-PARL databases used for acoustic model training, and the EUROPARL corpus [101] that amount to about 44 million of word occurrences. The resulting dictionaries contained 24 K words for the MAVIR development and test data processing, and 136 K words for the EPIC test data processing. An unknown word has also been added to the dictionary in order to deal with the speech segments that contain the lowest-occurrence words in the training data. The corresponding acoustic model for this unknown word is a single phone trained with the lowest occurrence words of the training speech data. As for the phone-based speech recognizers, an 8-g phone LM trained from the same counterpart sources than the word-based LM was employed for MAVIR development and test data. For EPIC test data, a 6-g phone LM trained from the same transcription data than the word-based LM for this same dataset has been employed. LM training has been conducted with the SRILM toolkit [98].

To process the speech data, this system first segmented the speech signals into more manageable chunks (which spread between 5 and 30 s), using a ITU-T G.729 VAD implementation. Then, each chunk is processed by the word and phone-based speech recognizers to obtain the corresponding set of words/phones recognized along with their confidence scores. Next, post-processing of the segments that contain words whose confidence score falls below a threshold (type 1) or the unknown word was recognized (type 2) was carried out. This threshold has been tuned on development data.

For type 1 segments, the corresponding speech segment is given to a GMM-HMM word matcher. This GMM-HMM word matcher takes the phone

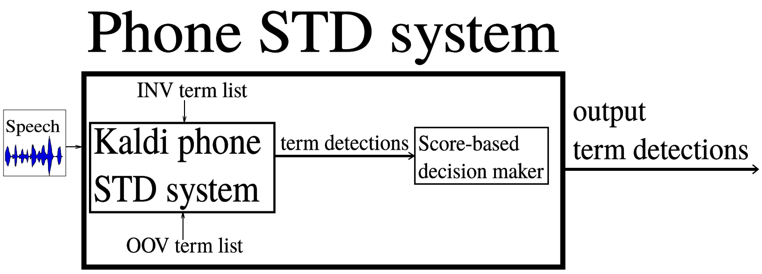

Fig. 6 Phone-based STD system architecture

transcription given by the phone based speech recognizer in the time intervals corresponding to the start and end times of the segment, and produces the most likely word/s given this phone transcription. To do so, a dictionary composed of the 5000 most frequent words in the training data plus the OOV terms is given. This GMM-HMM word matcher has been trained from a set of training data and the decoding of these speech data to learn from the phone recognition errors. These training data differ again for each data set. For MAVIR development and test data processing, the training data include the training data provided by the organizers. For EPIC test data processing, all the TC-STAR training data were employed. In the GMM-HMM word matcher, each phone is modeled with three emitting states, with a total number of Gaussians of 15,000.

For type 2 segments, an OOV term is expected to be found. To do so, a Levenshtein distance-based word identifier computes the Levenshtein distance between the phone sequence given by the phone-based recognizer and the phone transcription of the OOV term so that a low Levenshtein distance match suggests the occurrence of this OOV term. The threshold set for these type 2 segments has been tuned on development data.

Finally, the system merges the putative detections that come from three different sources: (1) All the terms detected by the word-based LVCSR system, (2) a single-word term output by the GMM-HMM word matcher that coincides with a term (for multi-word terms a new detection emerges in case all the words of the term appear in the output of the word-based LVCSR system or the GMM-HMM matcher) for type 1 segments, and (3) the OOV detections corresponding to the type 2 segments. The score given to each detection depends on the specific source. In this way, term detections from the word-based LVCSR system maintain the confidence score given by the recognition process; detections from the GMM-HMM word matcher take as score the confidence computed during the GMM-HMM word matching; and detections from the Levenshtein distance-based word identifier take a score in the opposite direction to the computed distance.

3.7 Phone-based STD system (phone STD)

This system bases on phone speech recognition and its architecture is shown in Fig. 6.

The ASR conducts phone-based speech recognition to produce phone lattices from the speech data. As features, the widely used 13-dimensional MFCCs with first and second order time derivatives and CMVN applied were first extracted. Following the steps of the Kaldi Librispeech s5 recipe, LDA is applied to reduce the feature dimensionality, which is then followed by MLLT to match the GMM diagonal assumption. Finally, to improve model robustness fMLLR and SAT are also applied. Regarding to acoustic model training, context-dependent GMM-HMMs, with 16 GMM components each, have been trained. As LM, 5-g phone LM has been trained with the SRILM toolkit [98]. Twenty-seven phones were modeled in this system.

For acoustic model training, the training speech data comprise different speech data sets:

- ALBAYZIN [104], which consists of phonetically balanced sentences that cover a wide range of variability in terms of speaker-dependent and phonetic factors, a collection of semantically and syntactically constrained sentences extracted from a geographic database query task, and a set of frequently used words and sentences recorded in clean and noisy environments (12:7 h in total).
- DOMOLAB [105], which corresponds to a corpus recorded in the kitchen of a home automation scenario, and contains a combination of speaker utterances for the automatic control of appliances where different acoustic environments and speaker locations are considered (9:2 h in total).
- SpeechDat-Car [106], which comprises a collection of speech resources to support training and testing of multilingual speech recognition applications in car environments (18:7 h in total).
- TC-STAR [107], which focuses at the translation of unconstrained conversational speech as it appeared in the broadcast (parliamentary) speeches and meetings from 2004 until 2007 (58:8 h in total).
- TDTDB, which corresponds to television (TV) program broadcast by the Spanish public TV (RTVE) during 2014, and includes a collection of audio recordings of multigenre data automatically transcribed manually validated (14:1 h in total).

In addition, the training data provided by the organizers (4 h in total) were also employed. Then, about 120 h were used for acoustic model training.

For language model training, about one million of word occurrences extracted from the Europarl corpus [101] were employed to train the 5-g phone LM.

Kaldi term detector [52–54] was employed to produce putative detections from the phone lattices given by the ASR subsystem and the corresponding canonical phone transcription for each term, and the decision maker outputs as detections those whose score derived from the Kaldi term detector is above a predefined threshold.

3.8 Mismatch-based phone STD system (mismatch phone STD)

This system is the same as the Phone STD system explained before with some term-based modifications aiming to increase term detection: Instead of considering the canonical phone sequence as phone transcription for each search term, multiple term pronunciations are considered to form proxy terms. To do so, a mismatch of one phone was considered to build the multiple term pronunciations (and hence the proxy terms). To keep the number of proxy terms reasonable, only the pronunciations whose occurrence probability obtained from a phone confusion matrix is above a predefined threshold were considered. This phone confusion matrix was trained on the training data employed for acoustic model training by running a phone recognition experiment. The threshold was estimated from the development data provided by the organizers.

3.9 Confusion matrix-based phone STD system (CM phone STD)

This system is the Phone STD system explained before with the use of a confusion matrix during term search. The idea is to expand the term search graph considering the whole ensemble of proxy terms. For this purpose, the mechanism of the Kaldi tool to generate proxy words [53] was employed, but deploying the term search at phone-level (instead of word-level) as referred in Eq. 3:

$$K^{'} = Project(ShortestPath(K \circ L \circ E)), \qquad (3)$$

where K is a finite-state acceptor for a certain OOV term, L is a finite-state transducer for the pronunciation of that OOV term, and E is an edit-distance transducer that maps the costs of alternative phone sequences given by a phone confusion matrix. This confusion matrix is the same as that used in the Mismatch Phone STD system explained before.

When several occurrences of the same term are output at the same time, the one with the highest score is kept.

3.10 Confusion matrix-based phone STD system + term-specific threshold (CM phone STD + TST)

This system aims to improve the decision maker of the CM Phone STD system explained before. To do so, instead of using a fixed threshold for selecting the most reliable detections, the TST approach [23] explained before has been employed to produce a term-dependent score for each detection.

4 Results and discussion

System results are presented in Table 9 for development data and Tables 10 and 11 for MAVIR test data and EPIC test data, respectively. The result ranking for development and test data for MAVIR database shows a different behavior. For development data, Syn-Syll Comb system performs the best, whereas for test data, Combined Kaldi system gets the best performance. This discrepancy is due to the Combined Kaldi system has a larger vocabulary than the Syn-Syll Comb system, which derives in a smaller out-of-vocabulary rate (defined as the number of terms in the MAVIR INV test term list that do not appear in the LVCSR system vocabulary), as shown in Table 12. The best performance of both systems on each set of data (development and test) compared to the rest of the systems is due to the more robust ASR subsystem (word-based ASR with acoustic model and LM trained on a great variety of speech and text sources). On development data, the use of a phone-based ASR subsystem in the Phone STD, Mismatch Phone STD, CM Phone STD, and CM, Phone STD + TST systems is clearly giving the worst overall performance compared with the other systems (that employ word-based ASR subsystems). However, these phone-based systems typically have a great advantage in terms of fast search and indexing, and the possibility of retrieving OOV terms without any additional system development. The performance on development data of the Syn-Syll Comb system is significantly better for a paired t-test compared with the other systems ($p < 0.01$) except with the Combined Kaldi and Proxy Kaldi systems, whose performance gaps are insignificant for a paired t test. The Kaldi matcher system obtains better performance than the DD Kaldi and DI

Table 9 System results of the ALBAYZIN STD 2016 evaluation on development data

System ID	MTWV	ATWV	p(FA)	p(Miss)
Combined Kaldi	0.5597	0.5551	0.00008	0.359
Proxy Kaldi	0.5155	0.5151	0.00015	0.334
Syn-Syll Comb	0.5729	0.5729	0.00009	0.337
DD Kaldi	0.2333	0.2315	0.00003	0.737
DI Kaldi	0.2118	0.2096	0.00004	0.748
Kaldi matcher	0.4828	0.4828	0.00006	0.458
Phone STD	0.0720	0.0613	0.00014	0.789
Mismatch Phone STD	0.0000	-2.5465	0.00000	1.000
CM Phone STD	0.1390	0.1358	0.00009	0.771
CM Phone STD+TST	0.1705	0.1545	0.00013	0.704

Table 10 System results of the ALBAYZIN STD 2016 evaluation on MAVIR test data

System ID	MTWV	ATWV	p(FA)	p(Miss)
Combined Kaldi	0.5850	0.5724	0.00008	0.336
Proxy Kaldi	0.5560	0.5506	0.00005	0.397
Syn-Syll Comb	0.5140	0.5090	0.00007	0.414
DD Kaldi	0.2165	0.2141	0.00006	0.721
DI Kaldi	0.2139	0.2122	0.00007	0.714
Kaldi matcher	0.4327	0.4311	0.00011	0.453
Phone STD	0.1323	0.1288	0.00013	0.734
Mismatch Phone STD	0.0000	-3.8487	0.00000	1.000
CM Phone STD	0.1816	0.1803	0.00011	0.710
CM Phone STD+TST	0.2098	0.2077	0.00010	0.691

Table 11 System results of the ALBAYZIN STD 2016 evaluation on EPIC test data

System ID	MTWV	ATWV	p(FA)	p(Miss)
Combined Kaldi	0.8416	0.8373	0.00005	0.105
Proxy Kaldi	0.8436	0.8388	0.00007	0.082
Syn-Syll Comb	0.8035	0.8023	0.00004	0.156
DD Kaldi	0.5507	0.5336	0.00007	0.378
DI Kaldi	0.5507	0.5376	0.00008	0.373
Kaldi matcher	0.8341	0.8296	0.00004	0.126
Phone STD	0.2637	0.2394	0.00028	0.452
Mismatch Phone STD	0.0000	-8.7347	0.00000	1.000
CM Phone STD	0.4028	0.3523	0.00019	0.412
CM Phone STD+TST	0.4354	0.4008	0.00019	0.372

Kaldi systems since the former incorporates more robust ASR subsystem (DNN-HMM as acoustic modeling and language model trained on a great variety of text sources) and STD subsystem. This better performance is significant for a paired *t* test ($p < 0.01$).

On MAVIR test data, the best performance corresponds to the Combined Kaldi system. This better performance is significant for a paired *t* test compared with the rest of the systems ($p < 0.01$) except with the Proxy Kaldi and Syn-Syll Comb systems. This is consistent with the findings in the development data. Again, Combined Kaldi, Proxy Kaldi, and Syn-Syll Comb systems outperform the rest due to the more robust ASR subsystem. In addition, these systems have smaller out-of-vocabulary rate than the rest, as shown in Table 12, which is clearly giving a better performance. As in the development data, the improvement of the Kaldi matcher system over the DD Kaldi and DI Kaldi systems is significant for a paired *t* test ($p < 0.01$) due to the more robust ASR and STD subsystems. An interesting finding is that the performance of the CM Phone STD + TST system, which bases on phone ASR, is quite near to that of the DD Kaldi and DI Kaldi systems, and that the performance gaps are insignificant for a paired *t* test. This confirms the power of phone-based ASR to build STD systems. Mostly, this insignificant performance gap is due to the OOV term retrieval capability of a phone-based STD system, which is partially absent in the DD Kaldi and DI Kaldi systems (see the OOV rate in Table 12), and the multi-term retrieval capability, which is absent in the DD Kaldi and DI Kaldi systems (see Section *Performance analysis of STD systems based on single/multi-word terms*). This confirms that a phone-based STD system can obtain a performance that is comparable to that of the word-based STD systems in case these do not incorporate a robust mechanism for OOV and multi-word term detection.

Table 12 Percentage of MAVIR INV test terms that do not appear in the LVCSR system vocabulary (only for word-based STD systems)

System ID	OOV rate
Combined Kaldi	7.3%
Proxy Kaldi	7.3%
Syn-Syll Comb	11.9%
DD Kaldi	19.8%
DI Kaldi	19.8%
Kaldi matcher	13.5%

On EPIC test data, the best performance is obtained by the Proxy Kaldi system. This performance is significant for a paired *t* test compared with all the systems ($p < 0.01$) except with the Syn-Syll Comb system whose performance gap is weak significant ($p < 0{:}04$), and with the Kaldi matcher and Combined Kaldi for which the performance gaps are insignificant. Again, the phone-based STD systems obtain the worst performance due to the absence of lexical information in the ASR subsystem. EPIC test data are *easier* data from an ASR point of view than the MAVIR data, since EPIC data contain clean speech. This is confirmed by the p.563 MOS values presented in Tables 2 and 3 for MAVIR and EPIC data, respectively, where the MOS values are generally better in the EPIC database. Therefore, the performance of all the systems for EPIC data is much better than that obtained on MAVIR data, whose acoustic conditions degrade the ASR performance, and hence the overall STD performance. In addition, as can be seen in Table 13, the lower OOV rate compared to MAVIR test data is also enhancing the final STD performance. On the other hand, the performance gaps between the word-based STD systems and the phone-based STD systems are larger than in MAVIR test data due to the clean speech condition. In addition, the better performance of the DD Kaldi and DI Kaldi systems compared with the phone-based STD systems is significant for a paired *t* test ($p < 0.01$). One interesting finding is the insignificant performance gap between the Proxy Kaldi and the Kaldi matcher systems. We consider this is due to the mismatch between the development data (noisy speech in general domain) and test data (clean speech in meeting domain), which may

Table 13 Percentage of EPIC INV test terms that do not appear in the LVCSR system vocabulary (only for word-based STD systems)

System ID	OOV rate
Combined Kaldi	0.6%
Proxy Kaldi	0.6%
Syn-Syll Comb	1.3%
DD Kaldi	7.3%
DI Kaldi	7.3%
Kaldi matcher	4.6%

degrade the performance of the best system due to the change in the application domain.

For all sets of data, the performance gap between MTWV and ATWV metrics is, in general, low. This means that the term detection scores are well-calibrated.

DET curves are shown in Figs. 7, 8, and 9 for MAVIR development data, MAVIR test data, and EPIC test data, respectively. On development data, the Proxy Kaldi system performs the best for low and high false alarm rates, whereas for moderate false alarm and miss rates, the Syn-Syll Comb system performs the best. The phone-based systems, as expected from the evaluation results, perform the worst in all the operation points. On MAVIR test data, the Proxy Kaldi system also performs the best for low and high false alarm rates, whereas for moderate false alarm and miss rates, the Combined Kaldi system performs the best. The CM Phone STD + TST and the CM Phone STD systems perform better than the DD Kaldi and DI Kaldi systems for low miss rate, and the reverse stands for low false alarm rate. On EPIC test data, the Proxy Kaldi system performs the best for almost all the operation points, and the performance gets near that of the Kaldi matcher for low miss rate. In general, the performance of the phone-based STD systems in EPIC data is the worst in almost all the operation points, except for low miss rate, where the DD Kaldi and DI Kaldi system performances are worse than that of three out of the four phone-based STD systems. All these results are consistent with the ATWV-based evaluation results.

4.1 Comparison to previous NIST STD evaluations

The results obtained in this ALBAYZIN STD 2016 evaluation cannot be directly compared with those of the NIST STD evaluations since 2006. However, some analysis aiming to provide some light across different languages/domains can be carried out. Comparing our results with those obtained in the first NIST STD evaluation held in 2006, it is clearly seen that better results are obtained in the ALBAYZIN STD 2016 evaluation compared to those obtained for the Arabic and Mandarin languages in the NIST STD 2006 evaluation, which agrees with the conclusions presented in [59]. However, for English language (which is *easier* than Arabic and Mandarin languages from an ASR perspective) on broadcast news and conversational telephone speech domains, for which typically enough training data exist, STD performance is near to that obtained in the EPIC data in our evaluation. These BNews and CTS domains are typically *easier* than the MAVIR domain, and hence better performance is expected. However, EPIC domain, which is much easier than MAVIR domain, obtains performance similar to the English language on BNews and CTS domains. Therefore, the difficulty of

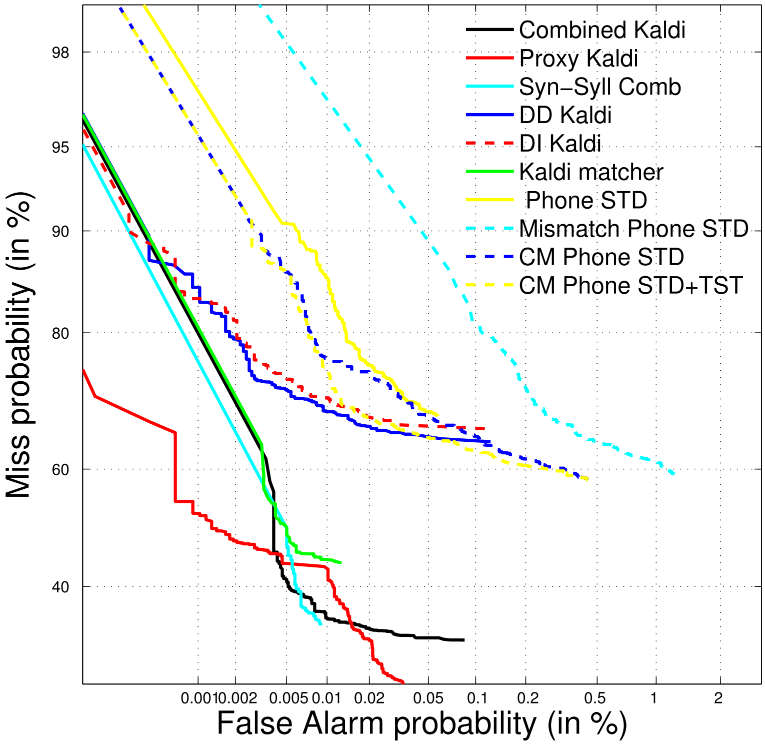

Fig. 7 DET curves of the STD systems for MAVIR development data

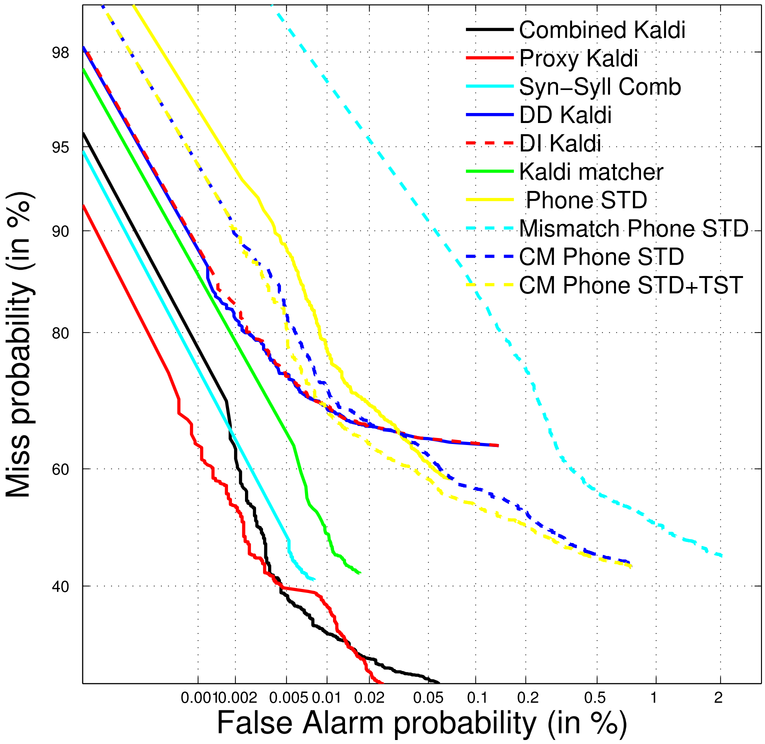

Fig. 8 DET curves of the STD systems for MAVIR test data

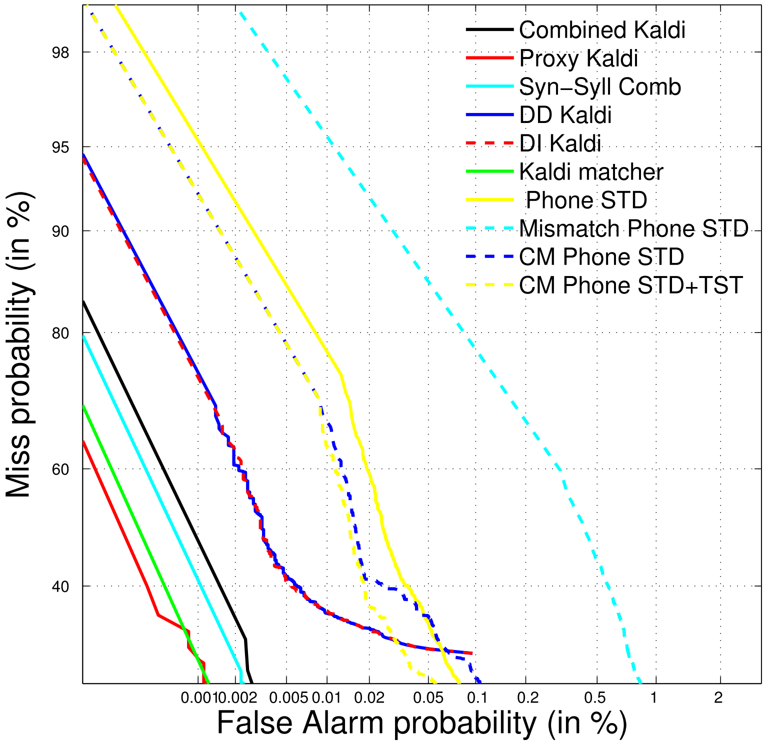

Fig. 9 DET curves of the STD systems for EPIC test data

BNews, CTS, and clean speech on parliament sessions can be considered *equivalent* for an STD perspective. However, when the diffculty of the domain increases (as is the case for meeting speech on English language), significant performance degradation is observed. It must be noted that the meeting speech employed in the English STD NIST 2006 evaluation contains more difficult speech than the parliament sessions corresponding to the EPIC data (e.g., the meeting speech contains more overlapping and noise than the parliament session speech). In addition, the progress of the STD technology since 2006 also contributes to enhance the system performance.

Comparing our results with those obtained in the NIST OpenKWS evaluations held from 2013 to 2016, similar performance is obtained on MAVIR data. Both in the NIST OpenKWS evaluations and ALBAYZIN STD evaluations, there is still enough room for improvement. On EPIC data, our results are much better than those obtained in the OpenKWS evaluations. This is expected, since EPIC data comprise a much *easier* domain than the MAVIR data for ASR, and hence from an STD perspective as well.

4.2 Performance analysis of STD systems based on term length

An analysis of the performance of the STD systems based on the length (in number of graphemes) of the test terms has been conducted and results are shown in Tables 14 and 15 for MAVIR and EPIC test data, respectively. Test terms have been divided into three categories: short-length terms (terms with up to 7 graphemes), medium-length terms (terms between 8 and 10 graphemes), and long-length terms (terms with more than 10 graphemes). In general, performance improves to a great extent from short to medium-length terms, since shorter words tend to produce more errors in ASR systems. However, for longer terms the improvement is not that clear and in most of the cases the STD performance for longer terms is worse than for medium-length terms. This may happen because, in word-based STD systems, longer terms are frequently composed of multiple words, which tend to decrease the STD performance since these words are more possible to convey ASR errors than a single word. For the phone-based STD systems presented, which are based on a sequence of phones, the term is simply treated as a sequence of phones (i.e., the number of words of the term does not play an important role). In this case, STD performance for long-length terms also degrades compared to medium-length terms, since they are more difficult to detect than medium-length terms (i.e., the sequence of phones output by the ASR subsystem has more ASR errors since more phones need to be *detected*). On the other hand, short-length terms typically cause so many FAs that the STD performance is also worse compared with medium-length terms.

Table 14 System results of the ALBAYZIN STD 2016 evaluation on MAVIR test data based on the term length (ATWV performance)

System ID	ATWV		
	Short	Medium	Long
Combined Kaldi	0.5239	0.6285	0.5346
Proxy Kaldi	0.5355	0.5546	0.5626
Syn-Syll Comb	0.4394	0.5644	0.4993
DD Kaldi	0.1920	0.2299	0.2141
DI Kaldi	0.1941	0.2246	0.2133
Kaldi matcher	0.3229	0.4759	0.4897
Phone STD	0.1274	0.1449	0.1019
Mismatch Phone STD	−3.9749	−5.0886	−1.4833
CM Phone STD	0.1603	0.2240	0.1280
CM Phone STD+TST	0.1793	0.2560	0.1582

'Short' denotes short-length terms (terms with up to 7 graphemes), 'Medium' denotes medium-length terms (terms between 8 and 10 graphemes), and 'Long' denotes long-length terms (terms with more than 10 graphemes)

4.3 Performance analysis of STD systems based on single/multi-word terms

An analysis of the performance of the STD systems for single-word and multi-word terms has been carried out and results are shown in Table 16. Results show performance degradation from single-word to multi-word terms for STD systems based on word units. This occurs because ASR errors affect term detection to a great extent when the terms are composed of more than one word. However, for STD systems based on phone recognition, this behavior is not that clear, since each term is simply treated as a sequence of phones, and hence the

Table 15 System results of the ALBAYZIN STD 2016 evaluation on EPIC test data based on the term length (ATWV performance)

System ID	ATWV		
	Short	Medium	Long
Combined Kaldi	0.8197	0.8504	0.8446
Proxy Kaldi	0.8076	0.8583	0.8611
Syn-Syll Comb	0.7848	0.8202	0.7973
DD Kaldi	0.5862	0.5023	0.4923
DI Kaldi	0.5992	0.5026	0.4854
Kaldi matcher	0.8229	0.8337	0.8342
Phone STD	0.1532	0.3005	0.2828
Mismatch Phone STD	−13.5174	−6.6105	−3.2178
CM Phone STD	0.2267	0.4507	0.3917
CM Phone STD+TST	0.3532	0.4498	0.3866

'Short' denotes short-length terms (terms with up to 7 graphemes), 'Medium' denotes medium-length terms (terms between 8 and 10 graphemes), and 'Long' denotes long-length terms (terms with more than 10 graphemes)

Table 16 System results of the ALBAYZIN STD 2016 evaluation on MAVIR test data for single-word and multi-word terms

System ID	Single			Multi		
	ATWV	p(FA)	p(Miss)	ATWV	p(FA)	p(Miss)
Combined Kaldi	0.6028	0.00009	0.303	0.2857	0.00000	0.452
Proxy Kaldi	0.5711	0.00005	0.364	0.3571	0.00000	0.405
Syn-Syll Comb	0.5276	0.00008	0.387	0.3333	0.00000	0.667
DD Kaldi	0.2368	0.00007	0.691	0.0000	0.00000	1.000
DI Kaldi	0.2347	0.00008	0.683	0.0000	0.00000	1.000
Kaldi matcher	0.4516	0.00013	0.420	0.2381	0.00000	0.762
Phone STD	0.1273	0.00012	0.745	0.1429	0.00000	0.857
Mismatch Phone STD	−4.2547	0.00000	1.000	−0.0208	0.00008	0.857
CM Phone STD	0.1868	0.00012	0.692	0.1190	0.00003	0.762
CM Phone STD+TST	0.1868	0.00012	0.692	0.2143	0.00003	0.762

'Single' refers to single-word terms and 'Multi' refers to multi-word terms

term length is more suitable than the single/multi-word term classification for further analysis. Results also show that multi-word term detection is still a challenging task in STD even for word-based STD systems.

4.4 Performance analysis of STD systems based on in-vocabulary/out-of-vocabulary terms

An analysis of the performance of the STD systems for in-vocabulary and out-of-vocabulary terms has been carried out and results are shown in Tables 17 and 18 for MAVIR and EPIC test data, respectively. Results show performance degradation from INV to OOV terms for all the word-based STD systems even though these incorporate some mechanism for OOV term retrieval. This suggests the difficulty of OOV term retrieval in STD systems no matter the target domain is *easy* as EPIC data, or *difficult* as MAVIR data. This performance degradation is due to the impossibility of using lexical information. On the other hand, performance is rather similar for INV and OOV terms (and even better for OOV terms in some cases) for the phone-based STD systems because for these systems INV and OOV are the *same* from the term retrieval point of view, since they do not use words. Since all the STD systems have to resort to phone based ASR or some other phone-based approach for OOV term retrieval all the differences between word, and phone-based STD become to a minimum when retrieving OOV terms, and hence many systems obtain *equivalent* OOV term performance.

4.5 Performance analysis of STD systems based on in-language/out-of-language terms

An analysis of the performance of the STD systems for in-language and out-of-language terms has been carried out and results are shown in Table 19. Performance degradation is observed from in-language to out-of-language terms. Whereas INL terms are terms whose pronunciation matches the target language, and for which enough data are typically employed to train both the acoustic models and LMs, for OOL terms, the pronunciation cannot be effectively derived from grapheme-to-phoneme rules given

Table 17 System results of the ALBAYZIN STD 2016 evaluation on MAVIR test data for INV and OOV terms

System ID	INV			OOV		
	ATWV	p(FA)	p(Miss)	ATWV	p(FA)	p(Miss)
Combined Kaldi	0.6324	0.00008	0.274	0.2084	0.00008	0.712
Proxy Kaldi	0.6404	0.00008	0.274	0.0059	0.00047	0.400
Syn-Syll Comb	0.5767	0.00005	0.365	0.0984	0.00014	0.754
DD Kaldi	0.2292	0.00007	0.697	0.1225	0.00002	0.852
DI Kaldi	0.2268	0.00008	0.690	0.1236	0.00002	0.849
Kaldi matcher	0.4791	0.00012	0.397	0.1399	0.00007	0.789
Phone STD	0.1320	0.00015	0.715	0.1090	0.00008	0.787
Mismatch Phone STD	−3.9621	0.00000	1.0000	−3.1612	0.00000	1.0000
CM Phone STD	0.1787	0.00012	0.702	0.1900	0.00005	0.755
CM Phone STD+TST	0.1787	0.00012	0.702	0.1972	0.00005	0.755

'INV' refers to in-vocabulary terms and 'OOV' refers to out-of-vocabulary terms

Table 18 System results of the ALBAYZIN STD 2016 evaluation on EPIC test data for INV and OOV terms

System ID	INV			OOV		
	ATWV	p(FA)	p(Miss)	ATWV	p(FA)	p(Miss)
Combined Kaldi	0.9290	0.00004	0.026	0.4206	0.00012	0.461
Proxy Kaldi	0.9309	0.00004	0.026	0.4203	0.00024	0.187
Syn-Syll Comb	0.9362	0.00003	0.035	0.1936	0.00010	0.709
DD Kaldi	0.5894	0.00008	0.311	0.2798	0.00001	0.669
DI Kaldi	0.5928	0.00007	0.324	0.2865	0.00001	0.669
Kaldi matcher	0.9268	0.00004	0.033	0.3877	0.00001	0.594
Phone STD	0.2324	0.00029	0.437	0.2714	0.00029	0.393
Mismatch Phone STD	−8.8765	0.00000	1.000	−8.0902	0.00000	1.000
CM Phone STD	0.3506	0.00020	0.386	0.3597	0.00009	0.528
CM Phone STD+TST	0.3506	0.00020	0.386	0.3617	0.00009	0.528

'INV' refers to in-vocabulary terms and 'OOV' refers to out-of-vocabulary terms

the target language, and are typically scarce in the acoustic model and LM training data.

It must be noted that the lowest performance degradation from INL terms to OOL terms corresponds to the DD Kaldi and DI Kaldi systems. This might be explained by the presence of some English terms in the dictionary of both systems.

4.6 Lessons learned

The ALBAYZIN Spoken Term Detection 2016 evaluation is integrated into a more general search-on-speech ALBAYZIN evaluation. This is the third edition of the search-on-speech ALBAYZIN evaluation, after those held in 2012 and 2014. This evaluation involves two different applications: Spoken Term Detection and Query-by-Example Spoken Term Detection. In the search-on-speech ALBAYZIN evaluation, the STD evaluation constitutes the third STD evaluation in Spanish language, after those held in 2012, and 2014. From the first evaluation, considerable improvements have been carried out in the evaluation organization aiming to offer an appealing evaluation for potential participants. In the first evaluation, only MAVIR data were employed. In addition, the list of terms only included single-word, INV, and INL terms. The second edition incorporated more difficult terms, since multi-word, and OOL terms were added on the same MAVIR data. In this third evaluation, the evaluation complexity is higher since two different domains were considered (MAVIR and EPIC data), and a real OOV term set was provided to participants. This means that participants were required to manage OOV term detection using an approach different to the traditional search on word lattices. In addition, some auxiliary data such as the word lattices provided by a Kaldi-based ASR system were also provided, but in the end none of the participants used them.

In previous evaluations, a single dataset for training/development was provided to participants, who may use the same data for training and development. This makes

Table 19 System results of the ALBAYZIN STD 2016 evaluation on MAVIR test data for in-language and out-of-language (foreign) terms

System ID	INL			OOL		
	ATWV	p(FA)	p(Miss)	ATWV	p(FA)	p(Miss)
Combined Kaldi	0.5985	0.00009	0.305	0.2167	0.00000	0.750
Proxy Kaldi	0.5719	0.00005	0.369	0.2609	0.00013	0.567
Syn-Syll Comb	0.5447	0.00008	0.374	0.0225	0.00002	0.950
DD Kaldi	0.2202	0.00007	0.712	0.1300	0.00001	0.833
DI Kaldi	0.2183	0.00007	0.705	0.1300	0.00001	0.833
Kaldi matcher	0.4453	0.00012	0.435	0.2381	0.00006	0.698
Phone STD	0.1382	0.00014	0.715	0.0000	0.00000	1.0000
Mismatch Phone STD	−4.1236	0.00000	1.0000	−0.1101	0.00000	1.0000
CM Phone STD	0.1935	0.00012	0.689	0.0000	0.00006	0.919
CM Phone STD+TST	0.1935	0.00012	0.689	−0.0954	0.00006	0.919

'INL' refers to Spanish terms and 'OOL' refers to foreign terms a fair comparison of the system performance on development data almost impossible. To solve this issue, in this evaluation the training/development speech dataset was explicitly divided into two different datasets (training and development) so that a meaningful analysis can be done on the development data.

This is the third time that MAVIR data have been employed in this round of ALBAYZIN STD evaluations. Although MAVIR data become very repetitive in the evaluations, we plan to use them again in next evaluations for comparison purposes aiming to evaluate the progress of the STD technology in Spanish. Regarding EPIC data, this is the first use of these data in the ALBAYZIN STD evaluation. Using two different domains is a straightforward way to compare system performances across two different domains, and has allowed us to examine the performance degradation of the systems depending on the nature of the speech data.

The ALBAYZIN STD evaluations held so far focused on finding disjoint term lists into disjoint speech data. In future evaluations, cross-search (searching development terms into test speech data, and searching test terms into development speech data) should also be taken into account. This will measure the generalization capability of the systems when searching for *known* terms within *unknown* speech data.

5 Conclusions

This paper has presented a spoken term detection international open evaluation for search-on speech in Spanish. The amount of systems submitted to the evaluation has made it possible to compare the progress of this technology under a common framework. System design and results along with a deep result analysis across different term characteristics (term length, INV/OOV, single/multi-word terms, and INL/OOL) have been presented. Five different research groups have taken part in the evaluation and ten different systems were submitted in total. All the submitted systems allow INV and OOV term detection. Some systems are based on phone ASR to retrieve OOV terms whereas others employ word lattices output by a word-based ASR system to produce OOV term detections. In addition, a full phone-based STD system has also been submitted. This phone-based STD system is suitable for fast indexing and search, although the performance is not as good as the rest of the systems based on word ASR. Given the challenge of the MAVIR data, the best performance can be considered high (ATWV = 0.5850). This performance is higher than that obtained in the previous ALBAYZIN STD 2014 evaluation (ATWV = 0.5350), which confirms the progress of the STD technology in Spanish, even though this year OOV terms have been explicitly designed, and the term list is more difficult (e.g., more OOL and multi-word terms). Regarding domain comparison, we have shown that for an easier domain such as that of the EPIC data with an easier term list (INV, INL, and single-word terms) performance is much better (ATWV = 0.8436).

We have also shown that OOV term detection still remains an important challenge in STD, as is the case with OOL and multi-word terms. For word-based STD systems longer words contribute to enhance the STD performance, and shorter words contribute to get lower the STD performance.

Given the best result obtained in the MAVIR data, there is still ample room for improvement. Due to the low rates obtained in OOV term detection on these data, the upcoming evaluations should focus on OOV terms aiming to encourage participants to build robust systems for OOV term detection. In addition, more OOL and multi-word terms could also be considered in the term list in future evaluations. All these results encourage us to maintain this evaluation in the future, trying to focus more on the challenges remaining in STD.

6 Endnotes

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

[2] http://www.isca-speech.org/iscaweb/index.php/sigs?layout=edit&id=132

[3] http://catalog.elra.info/product_info.php?products_id=1145

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

[5] http://cartago.lllf.uam.es/mavir/index.pl?m=videos

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

[7] http://www.itl.nist.gov/iad/mig/tests/std/2006

[8] http://www.tc-star.org

[9] http://cartago.lllf.uam.es/mavir/index.pl?m=descargas

[10] This database is presently being developed by the Software Technology Working Group (GTTS) research group of the University of the Basque Country (UPV/EHU), contact german.bordel@ehu.eus

Acknowledgements

This work has been partially funded by the Galician Government ('Centro singular de investigación de Galicia' accreditation 2016-2019 ED431G/01), the European Regional Development Fund (ERDF), the Spanish Ministry of Economy and Competitiveness (RESTORE project, TEC2015-67163-C2-1-R, ASLP-MULAN TIN2014-54288-C4-1-R, MINECO/FEDER,UE), and the Basque Government (BerbaOla project, KK-2017/00043). This work has also been partially funded by the project DSSL: Redes Profundas y Modelos de Subespacios para Deteccion y Seguimiento de Locutor Idioma y Enfermedades Degenerativas a partir de la Voz (TEC2015-68172-C2-1-P) founded by the Spanish Ministry of Economy and Competitiveness.

Author's contributions

JT and DTT designed and prepared the STD evaluation. They also built the DD Kaldi and DI Kaldi systems, and carried out the detailed analysis of the evaluation results presented in this paper. PLO and LDF built the Combined Kaldi and Proxy Kaldi systems. LS and IH built the Syn-Syll Comb system. ACL and JF built the Kaldi matcher system. JO and JL built the Phone STD, Mismatch Phone STD, CM Phone STD, and CM Phone STD + TST systems. The main contributions of this paper are as follows: systems submitted to the third spoken term detection evaluation for Spanish language are presented. Increasing complexity in the term list from previous Spoken Term Detection evaluations. Analysis of system results from two different domains is presented. Lessons learned from the STD evaluation are presented. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Escuela Politécnica Superior, Fundación Universitaria San Pablo CEU, Campus de Montepríncipe, Madrid, Spain. [2]AuDIas, Universidad Autónoma de Madrid, Av. Francisco Tomás y Valiente, 11. Escuela Politécnica Superior, Madrid, Spain. [3]Universidade da Coruña, IRLab, CITIC, Campus de Elviña s/n, A Coruña, Spain. [4]Multimedia Technologies Group (GTM), AtlantTIC Research Center, E. E. Telecomunicación, Campus Universitario de Vigo, s/n, Vigo, Spain. [5]Aholab (UPV/EHU), ETSI Bilbao, Alda. Urquijo s/n, Bilbao, Spain. [6]Speech Technology Group, Universidad Politécnica de Madrid, Ciudad Universitaria s/n, Madrid, Spain. [7]ViVoLab, Aragon Institute for Engineering Research (I3A) Universidad de Zaragoza, María de Luna 1, Zaragoza, Spain.

Received: 6 July 2017 Accepted: 12 September 2017
Published online: 29 September 2017

References

1. Larson, M., & Jones, G. (2011). Spoken content retrieval: A survey of techniques and technologies. *Foundations and Trends in Information Retrieval, 5*(4-5), 235–422.
2. Mamou, J., Ramabhadran, B., & Siohan, O. (2007). Vocabulary independent spoken term detection. In *Proc. of ACM SIGIR* (pp. 615–622).
3. Mamou, J., & Ramabhadran, B. (2008). Phonetic query expansion for spoken document retrieval. In *Proc. of Interspeech* (pp. 2106–2109).
4. Can, D., Cooper, E., Sethy, A., White, C., Ramabhadran, B., & Saraclar, M. (2009). Effect of pronunciations on OOV queries in spoken term detection. In *Proc. of ICASSP* (pp. 3957–3960).
5. Fiscus, J. G., Ajot, J., Garofolo, J. S., & Doddingtion, G. (2007). Results of the 2006 spoken term detection evaluation. In *Proc. of workshop on searching spontaneous conversational speech* (pp. 45–50).
6. Vergyri, D., Stolcke, A., Gadde, R. R., & Wang, W. (2006). The SRI 2006 spoken term detection system. In *Proc. of NIST spoken term detection workshop (STD 2006)* (pp. 1–15).
7. Vergyri, D., Shafran, I., Stolcke, A., Gadde, R. R., Akbacak, M., Roark, B., & Wang, W. (2007). The SRI/OGI 2006 spoken term detection system. In *Proc. of Interspeech* (pp. 2393–2396).
8. Akbacak, M., Vergyri, D., & Stolcke, A. (2008). Open-vocabulary spoken term detection using graphone-based hybrid recognition systems. *Proc. of ICASSP*, 5240–5243.
9. Szoke, I., Fapso, M., Karafiat, M., Burget, L., Grezl, F., Schwarz, P., Glembek, O., Matejka, P., Kopecky, J., & Cernocky, J. (2008). Spoken term detection system based on combination of LVCSR and phonetic search. In *Machine learning for multimodal interaction vol. 4892/2008* (pp. 237–247).
10. Szoke, I., Burget, L., Cernocky, J., & Fapso, M. (2008). Sub-word modeling of out of vocabulary words in spoken term detection. In *Proc. of SLT* (pp. 273–276).
11. Szoke, I., Fapso, M., Burget, L., & Cernocky, J. (2008). Hybrid word-subword decoding for spoken term detection. In *Proc. of speech search workshop at SIGIR* (pp. 42–48).
12. Meng, S., Yu, P., Liu, J., & Seide, F. (2008). Fusing multiple systems into a compact lattice index for Chinese spoken term detection. In *Proc. of ICASSP* (pp. 4345–4348).
13. Thambiratmann, K., & Sridharan, S. (2007). Rapid yet accurate speech indexing using dynamic match lattice spotting. *IEEE Trans. Audio Speech Lang. Process., 15*(1), 346–357.
14. Wallace, R., Vogt, R., Baker, B., & Sridharan, S. (2010). Optimising figure of merit for phonetic spoken term detection. In *Proc. of ICASSP* (pp. 5298–5301).
15. Parada, C., Sethy, A., Dredze, M., & Jelinek, F. (2010). A spoken term detection framework for recovering out-of-vocabulary words using the web. In *Proc. of Interspeech* (pp. 1269–1272).
16. Jansen, A., Church, K., & Hermansky, H. (2010). Towards spoken term discovery at scale with zero resources. In *Proc. of Interspeech* (pp. 1676–1679).
17. Parada, C., Sethy, A., & Ramabhadran, B. (2010). Balancing false alarms and hits in spoken term detection. In *Proc. of ICASSP* (pp. 5286–5289).
18. Schneider, D., Mertens, T., Larson, M., & Kohler, J. (2010). Contextual verification for open vocabulary spoken term detection. In *Proc. of Interspeech* (pp. 697–700).
19. Chan, C.-A., & Lee, L.-S. (2010). Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping. In *Proc. of Interspeech* (pp. 693–696).
20. Chen, C.-P., Lee, H.-Y., Yeh, C.-F., Lee, L.-S.: Improved spoken term detection by feature space pseudo-relevance feedback. In: Proc. of Interspeech, pp. 1672-1675 (2010).
21. Motlicek, P., Valente, F., & Garner, P. (2010). English spoken term detection in multilingual recordings. In *Proc. of Interspeech* (pp. 206–209).
22. Szoke, I., Fapso, M., Karafiat, M., Burget, L., Grezl, F., Schwarz, P., Glembek, O., Matejka, P., Kontar, S., & Cernocky, J. (2006). BUT system for NIST STD 2006 - English. In *Proc. of NIST spoken term detection evaluation workshop (STD'06)* (pp. 1–15).
23. Miller, D. R. H., Kleber, M., Kao, C.-L., Kimball, O., Colthurst, T., Lowe, S. A., Schwartz, R. M., & Gish, H. (2007). Rapid and accurate spoken term detection. In *Proc. of Interspeech* (pp. 314–317).
24. Li, H., Han, J., Zheng, T., & Zheng, G. (2012). A novel confidence measure based on context consistency for spoken term detection. In *Proc. of Interspeech* (pp. 2430–2433).
25. Lee, H.-Y., & Lee, L.-S. (2013). Enhanced spoken term detection using support vector machines and weighted pseudo examples. *IEEE Trans. Audio Speech Lang. Process., 21*(6), 1272–1284.
26. Chiu, J., & Rudnicky, A. (2013). Using conversational word bursts in spoken term detection. In *Proc. of Interspeech* (pp. 2247–2251).
27. Seide, F., Yu, P., Ma, C., & Chang, E. (2004). Vocabulary-independent search in spontaneous speech. In *Proc. of ICASSP* (pp. 253–256).
28. Lee, S., Tanaka, K., & Itoh, Y. (2015). Combination of diverse subword units in spoken term detection. In *Proceedings of Interspeech* (pp. 3685–3689).
29. Logan, B., Thong, J.-M. V., & Moreno, P. J. (2005). Approaches to reduce the effects of OOV queries on indexed spoken audio. *IEEE Transactions on Multimedia, 7*(5), 899–906.
30. Logan, B., Moreno, P., & Deshmuk, O. (2002). Word and sub-word indexing approaches for reducing the effects of OOV queries on spoken audio. In *Proc. of HLT* (pp. 31–35).
31. Ma, B., & Li, H. (2005). A phonotactic-semantic paradigm for automatic spoken document classification. In *Proc. of ACM SIGIR* (pp. 369–376).
32. Pinto, J., Szoke, I., Prasanna, S. R. M., & Hermansky, H. (2008). Fast approximate spoken term detection from sequence of phonemes. In *Proc. of ACM SIGIR* (pp. 28–33).
33. Ohno, T., & Akiba, T. (2013). DTW-distance-ordered spoken term detection. In *Proc. of Interspeech* (pp. 3737–3741).
34. Nakagawa, S., Iwami, K., Fujii, Y., & Yamamoto, K. (2013). A robust/fast spoken term detection method based on a syllable n-gram index with a distance metric. *Speech Comm., 55*(3), 470–485.
35. Su, H., Hieronymus, J., He, Y., Fosler-Lussier, E., & Wegmann, S. (2014). Syllable based keyword search: Transducing syllable lattices to word lattices. In *Proc. of SLT* (pp. 489–494).
36. Wang, D., Frankel, J., Tejedor, J., & King, S. (2008). A comparison of phone and grapheme-based spoken term detection. In *Proc. of ICASSP* (pp. 4969–4972).
37. Wallace, R., Vogt, R., & Sridharan, S. (2007). A phonetic search approach to the 2006 NIST spoken term detection evaluation. In *Proc. of Interspeech* (pp. 2385–2388).
38. Parlak, S., & Saraclar, M. (2008). Spoken term detection for Turkish broadcast news. In *Proc. of ICASSP* (pp. 5244–5247).
39. Kanda, N., Takeda, R., & Obuchi, Y. (2012). Using rhythmic features for japanese spoken term detection. In *Proc. of SLT* (pp. 170–175).
40. Wollmer, M., Schuller, B., & Rigoll, G. (2013). Keyword spotting exploiting long short-term memory. *Speech Comm., 55*(2), 252–265.
41. Wang, D. (2010). Out-of-Vocabulary Spoken Term Detection. PhD Thesis, University of Edinburgh.
42. James, D. A. (1994). A system for unrestricted topic retrieval from radio news broadcasts. In *Proc. of ICASSP* (pp. 279–282).
43. Jones, G. J. F., Foote, J. T., Sparck Jones, K., & Young, S. J. (1996). Retrieving spoken documents by combining multiple index sources. In *Proc. of ACM SIGIR* (pp. 30–38).
44. Saraclar, M., & Sproat, R. (2004). Lattice-based search for spoken utterance retrieval. In *Proc. of HLT-NAACL* (pp. 129–136).

45. Iwata, K., Shinoda, K., Furui, S.: Robust spoken term detection using combination of pone based and word-based recognition. In: Proc. of Interspeech, pp. 2195-2198 (2008).
46. Li, J., Wang, X., & Xu, B. (2014). An empirical study of multilingual and low-resource spoken term detection using deep neural networks. In *Proc. of Interspeech* (pp. 1747–1751).
47. Yu, P., & Seide, F. (2004). A hybrid word / phoneme-based approach for improved vocabulary independent search in spontaneous speech. In *Proc. of ICSLP* (pp. 293–296).
48. Yazgan, A., & Saraclar, M. (2004). Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition. In *Proc. of ICASSP* (pp. 745–748).
49. Xu, D., & Metze, F. (2014). Word-based probabilistic phonetic retrieval for low-resource spoken term detection. In *Proc. of Interspeech* (pp. 2774–2778).
50. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P. The HTK Book (v3. 4). Engineering Department, Cambridge University, (2009). Cambridge: Engineering Department, Cambridge University.
51. Lee, K.-F., Hon, H.-W., & Reddy, R. (1990). An overview of the SPHINX speech recognition system. *IEEE Trans. Acoust. Speech Signal Process., 38*(1), 35–45.
52. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The KALDI speech recognition toolkit. In *Proc. of ASRU*.
53. Chen, G., Khudanpur, S., Povey, D., Trmal, J., Yarowsky, D., & Yilmaz, O. (2013). Quantifying the value of pronunciation lexicons for keyword search in low resource languages. In *Proc. of ICASSP* (pp. 8560–8564).
54. Pham, V. T., Chen, N. F., Sivadas, S., Xu, H., Chen, I.-F., Ni, C., Chng, E. S., & Li, H. (2014). System and keyword dependent fusion for spoken term detection. In *Proc. of SLT* (pp. 430–435).
55. Chen, G., Yilmaz, O., Trmal, J., Povey, D., & Khudanpur, S. (2013). Using proxies for OOV keywords in the keyword search task. In *Proc. of ASRU* (pp. 416–421).
56. Taras, B., & Nadeu, C. (2011). Audio segmentation of broadcast news in the Albayzin-2010 evaluation: Overview, results, and discussion. *EURASIP Journal on Audio, Speech, and Music Processing, 1*, 1–10.
57. Zelenak, M., Schulz, H., & Hernando, J. (2012). Speaker diarization of broadcast news in Albayzin 2010 evaluation campaign. *EURASIP Journal on Audio, Speech, and Music Processing, 19*, 1–9.
58. Rodríguez-Fuentes, L. J., Penagarikano, M., Varona, A., Díez, M., & Bordel, G. (2011). The Albayzin 2010 language recognition evaluation. In *Proceedings of Interspeech* (pp. 1529–1532).
59. Tejedor, J., Toledano, D. T., Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C., Cardenal, A., Echeverry-Correa, J. D., Coucheiro-Limeres, A., Olcoz, J., & Miguel, A. (2015, 2015). Spoken term detection ALBAYZIN 2014 evaluation: Overview, systems, results, and discussion. *EURASIP, Journal on Audio, Speech and Music Processing*, (21), 21:1–21:27.
60. Tejedor, J., Toledano, D. T., Anguera, X., Varona, A., Hurtado, L. F., Miguel, A., & Colás, J. (2013, 2013). Query-by-example spoken term detection ALBAYZIN 2012 evaluation: Overview, systems, results, and discussion. *EURASIP, Journal on Audio, Speech and Music Processing*, (23), 23:1–23:17.
61. Tejedor, J., Toledano, D. T., Lopez-Otero, P., Docio-Fernandez, L., & Garcia-Mateo, C. (2016, 2016). Comparison of ALBAYZIN query-by-example spoken term detection 2012 and 2014 evaluations. EURASIP. *Journal on Audio, Speech and Music Processing*, (1), 1:1–1:19.
62. Castán, D., Tavarez, D., Lopez-Otero, P., Franco-Pedroso, J., Delgado, H., Navas, E., Docio-Fernández, L., Ramos, D., Serrano, J., Ortega, A., Lleida, E.: Albayzín-2014 evaluation: Audio segmentation and classification in broadcast news domains. EURASIP, journal on audio, speech and music processing 2015(33), 33:1-33:9 (2015).
63. Méndez, F., Docío, L., Arza, M., & Campillo, F. (2010). The Albayzin 2010 text-to-speech evaluation. In *Proceedings of FALA* (pp. 317–340).
64. Billa, J., Ma, K. W., & McDonough, J. W. (1997). *Zavaliagkos, miller, D.R., Ross, K.N., el-Jaroudi, a.: Multilingual speech recognition: The 1996 Byblos callhome system*. Proc. of Eurospeech: In.
65. Cuayahuitl, H., & Serridge, B. (2002). Out-of-vocabulary word modeling and rejection for Spanish keyword spotting systems. In *Proc. of MICAI* (pp. 156–165).
66. Killer, M., Stuker, S., & Schultz, T. (2003). Grapheme based speech recognition. In *Proc. of Eurospeech* (pp. 3141–3144).
67. Tejedor, J.: Contributions to keyword spotting and spoken term detection for information retrieval in audio mining. PhD thesis, Universidad Autónoma de Madrid, Madrid, Spain (2009).
68. Burget, L., Schwarz, P., Agarwal, M., Akyazi, P., Feng, K., Ghoshal, A., Glembek, O., Goel, N., Karafiat, M., Povey, D., Rastrow, A., Rose, R. C., & Thomas, S. (2010). Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models. In *Proc. of ICASSP* (pp. 4334–4337).
69. Tejedor, J., Toledano, D. T., Wang, D., King, S., & Colás, J. (2014). Feature analysis for discriminative confidence estimation in spoken term detection. *Computer Speech and Language, 28*(5), 1083–1114.
70. Tejedor, J., Colás, J. (2006). Spanish keyword spotting system based on filler models, pseudo n-gram language model and a confidence measure. In *Proc. of IV Jornadas en Tecnologías del Habla* (pp. 255-260).
71. Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). The DET curve in assessment of detection task performance. In *Proc. of Eurospeech* (pp. 1895–1898).
72. NIST: Evaluation Toolkit (STDEval) Software. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA (1996). National Institute of Standards and Technology (NIST). http://www.itl.nist.gov/iad/mig/tests/std/tools. Accessed 19 Sept 2017.
73. ITU-T Recommendation P.563: Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications. http://www.itu.int/rec/T-REC-P.563/en. Accessed 19 Sept 2017.
74. Karakos, D., Bulyko, I., Schwartz, R., Tsakalidis, S., Nguyen, L., & Makhoul, J. (2014). Normalization of phonetic keyword search scores. In *Proc. of ICASSP* (pp. 7834–7838).
75. Trmal, J., Chen, G., Povey, D., Khudanpur, S., Ghahremani, P., Zhang, X., Manohar, V., Liu, C., Jansen, A., Klakow, D., Yarowsky, D., & Metze, F. (2014). A keyword search system using open source software. In *Proc. of SLT* (pp. 530–535).
76. Zhang, L., Karakos, D., Hartmann, W., Hsiao, R., Schwartz, R., & Tsakalidis, S. (2015). Enhancing low resource keyword spotting with automatically retrieved web documents. In *Proceedings of Interspeech* (pp. 839–843).
77. Wang, H., Ragni, A., Gales, M. J. F., Knill, K. M., Woodland, P. C., & Zhang, C. (2015). Joint decoding of tandem and hybrid systems for improved keyword spotting on low resource languages. In *Proceedings of Interspeech* (pp. 3660–3664).
78. Lv, Z., Cai, M., Zhang, W.-Q., & Liu, J. (2016). A novel discriminative score calibration method for keyword search. In *Proceedings of Interspeech* (pp. 745–749).
79. Hartmann, W., Zhang, L., Barnes, K., Hsiao, R., Tsakalidis, S., & Schwartz, R. (2016). Comparison of multiple system combination techniques for keyword spotting. In *Proceedings of Interspeech* (pp. 1913–1917).
80. Chen, N. F., Pharri, V. T., Xu, H., Xiao, X., Do, V. H., Ni, C., Chen, I.-F., Sivadas, S., Lee, C. H., Chng, E. S., Ma, B., & Li, H. (2016). Exemplar-inspired strategies for low-resource spoken keyword search in Swahili. In *Proceedings of ICASSP* (pp. 6040–6044).
81. Ni, C., Leung, C.-C., Wang, L., Liu, H., Rao, F., Lu, L., Chen, N. F., Ma, B., & Li, H. (2016). Cross lingual deep neural network based submodular unbiased data selection for low-resource keyword search. In *Proceedings of ICASSP* (pp. 6015–6019).
82. Cai, M., Lv, Z., Lu, C., Kang, J., Hui, L., Zhang, Z., & Liu, J. (2015). High-performance swahili keyword search with very limited language pack: The THUEE system for the OpenKWS15 evaluation. In *Proceedings of ASRU* (pp. 215–222).
83. Chen, N. F., Ni, C., Chen, I.-F., Sivadas, S., Pham, V. T., Xu, H., Xiao, X., Lau, T. S., Leow, S. J., Lim, B. P., Leung, C.-C., Wang, L., Lee, C.-H., Goh, A., Chng, E. S., Ma, B., & Li, H. (2015). Low resource keyword search strategies for Tamil. In *Proceedings of ICASSP* (pp. 5366–5370).
84. Mangu, L., Saon, G., Picheny, M., & Kingsbury, B. (2015). Order-free spoken term detection. In *Proceedings of ICASSP* (pp. 5331–5335).
85. NIST: OpenKWS13 Keyword Search Evaluation Plan. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA (2013). National Institute of Standards and Technology (NIST). https://www.nist.gov/sites/default/files/documents/itl/iad/mig/OpenKWS13-evalplan-v4.pdf. Accessed 19 Sept 2017.
86. NIST: Draft KWS14 Keyword Search Evaluation Plan. National Institute of Standards and Technology(NIST), Gaithersburg, MD, USA (2013). National Institute of Standards and Technology (NIST). https://www.nist.gov/sites/default/files/documents/itl/iad/mig/KWS14-evalplan-v11.pdf. Accessed 19 Sept 2017.
87. NIST: KWS15 Keyword Search Evaluation Plan. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA (2015). National Institute of Standards and Technology (NIST). https://www.nist.gov/sites/default/files/documents/itl/iad/mig/KWS15-evalplan-v05.pdf. Accessed 19 Sept 2017.

88. NIST: Draft KWS16 Keyword Search Evaluation Plan. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA (2016). National Institute of Standards and Technology (NIST). https://www.nist.gov/sites/default/files/documents/itl/iad/mig/KWS16-evalplan-v04.pdf. Accessed 19 Sept 2017.
89. Akiba, T., Nishizaki, H., Aikawa, K., Kawahara, T., & Matsui, T. (2011). Overview of the IR for spoken documents task in NTCIR-9 workshop. In *Proceedings of NTCIR-9* (pp. 1–13).
90. Akiba, T., Nishizaki, H., Aikawa, K., Hu, X., Itoh, Y., Kawahara, T., Nakagawa, S., Nanjo, H., & Yamashita, Y. (2013). Overview of the NTCIR-10 SpokenDoc-2 task. In *Proceedings of NTCIR-10* (pp. 1–15).
91. Akiba, T., Nishizaki, H., Nanjo, H., & Jones, G. J. F. (2014). Overview of the NTCIR-11 SpokenQuery&doc Task. In *Proceedings of NTCIR-11* (pp. 1–15).
92. Akiba, T., Nishizaki, H., Nanjo, H., & Jones, G. J. F. (2016). Overview of the NTCIR-12 SpokenQuery&doc-2 task. In *Proceedings of NTCIR-12* (pp. 1–13).
93. Vesely, K., Ghoshal, A., Burget, L., & Povey, D. (2013). Sequence-discriminative training of deep neural networks. In *Proceedings of Interspeech* (pp. 2345–2349).
94. Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., & Khudanpur, S. (2014). A pitch extraction algorithm tuned for automatic speech recognition. In *Proceedings of ICASSP* (pp. 2494–2498).
95. Povey, D., Hannemann, M., Boulianne, G., Burget, L., Ghoshal, A., Janda, M., Karafiat, M., Kombrink, S., Motlicek, P., Qian, Y., Riedhammer, K., Vesely, K., & Vu, N. T. (2012). Generating exact lattices in the WFST framework. In *Proceedings of ICASSP* (pp. 4213–4216).
96. Garcia-Mateo, C., Dieguez-Tirado, J., Docio-Fernandez, L., & Cardenal-Lopez, A. (2004). Transcrigal: A bilingual system for automatic indexing of broadcast news. In *Proceedings of LREC*. In.
97. Moreno, A., & Campillos, L. (2004). MAVIR: A corpus of spontaneous formal speech in spanish and english. In *Proceedings of Iberspeech* (pp. 224–230).
98. Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In *Proc. of Interspeech* (pp. 901–904).
99. Can, D., & Saraclar, M. (2011). Lattice indexing for spoken term detection. *IEEE Trans. Audio Speech Lang. Process., 19*(8), 2338–2347.
100. Zhang, X., Trmal, J., Povey, D., & Khudanpur, S. (2014). Improving deep neural network acoustic models using generalized maxout networks. In *Proceedings of ICASSP* (pp. 215–219).
101. Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th machine translation summit* (pp. 79–86).
102. Sainz, I., Erro, D., Navas, E., Hernaez, I., Sanchez, J., & Saratxaga, I. (2012). Aholab speech synthesizer for ALBAYZIN 2012 speech synthesis evaluation. In *Proceedings of Iberspeech* (pp. 645–652).
103. Mostefa, D., Hamon, O., Moreau, N., & Choukri, K. (2007). Evaluation report for the technology and corpora for speech to speech translation. In *TC-STAR project. Deliverable N* (p. 30).
104. Casacuberta, F., Garcia, R., Llisterri, J., Nadeu, C., Pardo, J., & Rubio, A. (1991). Development of spanish corpora for speech research (ALBAYZIN). In *Proceedings of workshop on international cooperation and standardization of speech databases and speech I/O Assesment methods* (pp. 26–28).
105. Justo, R., Saz, O., Guijarrubia, V., Miguel, A., Torres, M. I., & Lleida, E. (2008). Improving dialogue systems in a home automation environment. In *Proceedings of the 1st international conference on ambient media and systems* (pp. 1–6).
106. Moreno, A., Lindberg, B., Draxler, C., Richard, G., Choukri, K., Euler, S., & Allen, J. (2000). Speechdat-car: A large speech database for automotive environments. In *Proceedings of LREC*.
107. den Heuvel, H. V., Choukri, K., Gollan, C., Moreno, A., & Mostefa, D. (2006). TC-STAR: New language resources for ASR and SLT purposes. In *Proceedings of LREC* (pp. 2570–2573).