Tejedor *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2019) 2019:16
https://doi.org/10.1186/s13636-019-0159-7

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

ALBAYZIN 2018 spoken term detection evaluation: a multi-domain international evaluation in Spanish

Javier Tejedor[1*], Doroteo T. Toledano[2], Paula Lopez-Otero[3], Laura Docio-Fernandez[4], Ana R. Montalvo[5], Jose M. Ramirez[5], Mikel Peñagarikano[6] and Luis Javier Rodriguez-Fuentes[6]

Abstract

Search on speech (SoS) is a challenging area due to the huge amount of information stored in audio and video repositories. Spoken term detection (STD) is an SoS-related task aiming to retrieve data from a speech repository given a textual representation of a search term (which can include one or more words). This paper presents a multi-domain internationally open evaluation for STD in Spanish. The evaluation has been designed carefully so that several analyses of the main results can be carried out. The evaluation task aims at retrieving the speech files that contain the terms, providing their start and end times, and a score that reflects the confidence given to the detection. Three different Spanish speech databases that encompass different domains have been employed in the evaluation: the MAVIR database, which comprises a set of talks from workshops; the RTVE database, which includes broadcast news programs; and the COREMAH database, which contains 2-people spontaneous speech conversations about different topics. We present the evaluation itself, the three databases, the evaluation metric, the systems submitted to the evaluation, the results, and detailed post-evaluation analyses based on some term properties (within-vocabulary/out-of-vocabulary terms, single-word/multi-word terms, and native/foreign terms). Fusion results of the primary systems submitted to the evaluation are also presented. Three different research groups took part in the evaluation, and 11 different systems were submitted. The obtained results suggest that the STD task is still in progress and performance is highly sensitive to changes in the data domain.

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

1 Introduction

Search on speech (SoS) has become an interesting research area due to the huge amount of information stored in audio and video repositories. SoS focuses on retrieving speech content from audio repositories that matches user queries, for which the development of efficient methods is highly necessary [1]. Significant research has been carried out in SoS for spoken document retrieval (SDR) [2–7], keyword spotting (KWS) [8–13], spoken term detection (STD) [14–19], and query-by-example (QbE) STD and SDR [20–25]. STD is important due to the following factors: (1) It offers the possibility of retrieving any speech file that contains any term (a sequence of one or more words) from its textual representation, allowing search of any term in a large index efficiently. (2) This technology can be accessed using any device with text input capabilities. (3) It is suitable for building open-vocabulary SoS systems.

1.1 Spoken term detection overview

STD has been receiving much interest for years from outstanding companies/research institutes such as IBM [14, 26–30], BBN [31–33], SRI & OGI [34–36], BUT [17, 37, 38], Microsoft [39], QUT [40, 41], JHU [16, 42–44], Fraunhofer IAIS/NTNU/TUD [15], NTU [45, 46], and IDIAP [47], among others. STD systems are composed of two main stages: (1) indexing, which is usually done with an automatic speech recognition (ASR) subsystem and (2) search by a detection subsystem, as depicted

*Correspondence: javier.tejedornoguerales@ceu.es
[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

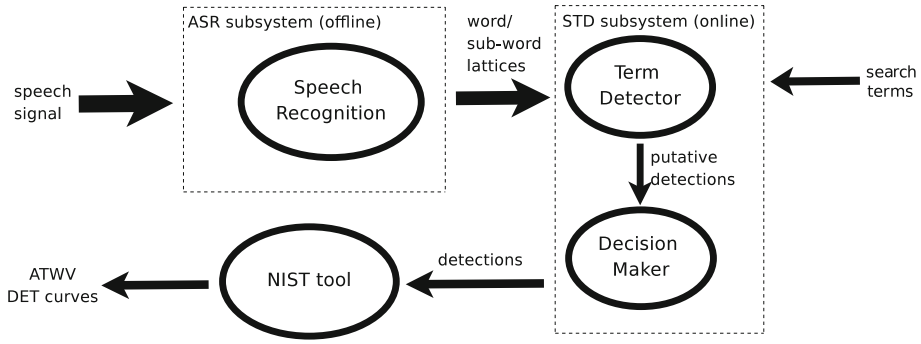

Fig. 1 Architecture and evaluation of a standard STD system

in Fig. 1. The ASR subsystem generates word/subword lattices from the input speech signal and stores them 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 *index*, and the decision maker decides whether each putative detection is a hit or a false alarm (FA) based on certain confidence measures.

For the ASR stage, word-based speech recognition has been widely used [35, 48–54], since this typically yields better performance than subword-based ASR [55–62] due to the lexical and language model (LM) information employed by the word-based ASR. However, one of the main drawbacks of word-based ASR is that it can only detect 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 vocabulary of the recognizer, i.e., out-of-vocabulary (OOV) terms. The combination of these two approaches has been proposed in order to exploit the relative advantages of word and subword-based strategies [17, 32, 33, 36, 44, 63–70].

Recently, end-to-end ASR-free approaches for STD have also been proposed, which aim to solve the issue of collecting and handling large amounts of data for building word and phone-based STD systems [28–30].

The availability of ASR tools, e.g., Hidden Markov Model Toolkit (HTK) [71], Sphinx [72], and Kaldi [44, 73], among others, facilitates the development of STD systems, since these mitigate the issue of constructing an ASR system from scratch. Among these, Kaldi is specially suitable for building STD systems since it integrates an ASR subsystem, a term detector, and a decision maker [73–75]. The Kaldi STD system employs a word-based approach for term detection, and a method based on proxy words (i.e., replace each OOV word by the most similar in-vocabulary word or word sequence) to detect OOV terms [76].

1.2 Methods

Research carried out in a certain area (speech recognition, speaker recognition, speaker diarization, to cite some examples) may be difficult to compare in the absence of a common evaluation framework. In STD, research also suffers from this issue since the published systems typically employ different acoustic databases and different lists of terms that make system comparison impossible. In this context, international evaluations provide a unique framework to measure the progress of any technology, as STD in this case.

ALBAYZIN evaluation campaigns comprise an internationally 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 biennially 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 [77–86].

Spanish is a major language in the world and significant research has been conducted on it for ASR, KWS, and STD tasks [87–93]. The increasing interest in SoS around the world and the lack of SoS evaluations dealing with the Spanish language encouraged us to organize a series of STD evaluations starting in 2012 and held biennially until 2018 aiming to evaluate the progress in this technology for Spanish. Each evaluation has been extended by incorporating new challenges. The main novelty of the fourth ALBAYZIN STD evaluation is the addition of a new data domain, namely broadcast news, with programs from the Spanish public television Radio Televisión Española (RTVE). In addition, a novel conversational speech database has also been used to assess the validity of the submitted systems in an unseen domain.

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

Moreover, the terms used in one of the databases in the ALBAYZIN 2016 STD evaluation were kept to enable a straightforward comparison of the systems submitted to both evaluations.

The main objectives of this evaluation can be summarized as follows:

- Organize the first Spanish STD multi-domain evaluation whose systems are ranked according to different databases and different domains
- Provide evaluation and benchmark with increasing complexity in the search terms compared to the previous ALBAYZIN STD evaluations

This evaluation is suitable for research groups/companies that work in speech recognition.

This paper is organized as follows: First, the Section 2 presents the evaluation (including databases, evaluation metrics, and participants) and a comparison with other STD evaluations. Then, in the Section 3, the different systems submitted to the evaluation are presented. Evaluation results along with a discussion are presented in the Section 4 which includes the corresponding paired t tests [94] as statistical significance measure for system comparison. The Section 5 presents a post-evaluation analysis based on some search term properties and the fusion of the primary systems submitted to the evaluation. The last section outlines the main conclusions of the paper.

2 Spoken term detection evaluation

2.1 STD evaluation overview

This evaluation involves searching a list of terms (given in written form) within speech data, and indicating the audio files and timestamps of each detected occurrence.

The evaluation consists in searching different term lists within different sets of speech data. Speech data comprise different domains (workshop talks, broadcast news, and 2-people conversations), for which individual datasets are given. Each domain contains training/development/test data, except the 2-people conversation dataset that only contains test data. The evaluation result ranking is based on the average system performance on the three datasets in the test experiment. Participants can use the training data for system training and the development data for system tuning, but any additional data can also be employed both for training and development.

Two different types of terms are defined in this evaluation, namely in-vocabulary terms and out-of-vocabulary terms. 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 4 contrastive systems. No manual intervention was allowed for each developed system to generate the final output file, and hence, all the systems had to be fully automatic. Listening to the test data, or any other human interaction with the test data, was forbidden before the evaluation results had been sent back to the participants. The output file with the term detections followed the standard Extensible Markup Language (XML)-based format accepted by the National Institute of Standards and Technology (NIST) evaluation tool [95]. Ground-truth labels corresponding to the test data were given to participants once the organizers sent 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 called a *false alarm*. If an actual occurrence is not detected, this is called a *miss*. The Actual Term-Weighted Value (ATWV) metric proposed by NIST [95] 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:

$$\mathrm{ATWV} = \frac{1}{|\Delta|} \sum_{K \in \Delta} \left(\frac{N_{\mathrm{hit}}^{K}}{N_{\mathrm{true}}^{K}} - \beta \frac{N_{\mathrm{FA}}^{K}}{T - N_{\mathrm{true}}^{K}} \right), \tag{1}$$

where Δ denotes the set of terms and $|\Delta|$ is the number of terms in this set. N_{hit}^{K} and N_{FA}^{K} represent the numbers of hits and false alarms of term K, respectively, and N_{true}^{K} is the number of actual occurrences of K in the audio. T denotes the audio length in seconds, and β is a weight factor set to 999.9, as in the ATWV proposed by NIST [31]. This weight factor causes an emphasis placed on recall compared to precision with a ratio 10:1.

ATWV represents the Term-Weighted Value (TWV) for a threshold given by the STD system (usually tuned on development data). An additional metric, called Maximum Term-Weighted Value (MTWV) [95] can also be used to evaluate the performance of an STD system. MTWV is the maximum TWV obtained by the STD system for all possible thresholds, and hence does not depend on the tuned threshold. Therefore, MTWV represents an upper-bound of the performance obtained by the STD system. Results based on this metric are also presented to evaluate system performance regardless the decision threshold.

p(Miss) and p(FA) values, which represent the probability of miss and FA of the STD system, respectively, are also reported. They are defined as follows:

$$\text{p(Miss)} = 1 - \frac{N_{\text{hit}}}{N_{\text{true}}} \tag{2}$$

$$\text{p(FA)} = \frac{N_{\text{FA}}}{T - N_{\text{true}}}, \tag{3}$$

where N_{hit} represents the number of hits of the STD system, N_{true} is the number of occurrences of the terms in the audio, N_{FA} represents the number of FAs of the STD system, and T denotes the audio length in seconds. These values provide a quantitative way to measure the STD system performance in terms of misses (or equivalently, hits) and false alarms.

In addition to ATWV, MTWV, p(Miss), and p(FA), NIST also proposed a detection error tradeoff (DET) curve [96] 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.

In this work, the NIST STD evaluation tool [97] was employed to compute MTWV, ATWV, p(Miss), p(FA), and DET curves.

2.3 Databases

Three different databases that comprise different acoustic conditions and domains have been employed for the evaluation. (1) For comparison purposes, the same MAVIR database employed in the previous ALBAYZIN STD evaluations in 2012, 2014, and 2016 has been used. (2) A database named RTVE that consists of different programs recorded from the Spanish public television (Radio Televisión Española) and involves different broadcast news domains. (3) The COREMAH database, which contains conversational speech with two speakers per recording. For the MAVIR and RTVE databases, three separate datasets (i.e., for training, development, and test) were provided to participants. For the COREMAH database, only test data were provided. This allowed measuring the generalization capability of the systems in an unseen domain. Tables 1, 2, and 3 include some database features such as the division into training, development, and test; the number of word occurrences; duration; and average mean opinion score (MOS) [98] as a way to get an idea of the quality of each speech file in the different databases.

2.3.1 MAVIR

The MAVIR database consists of a set of Spanish talks extracted from the MAVIR workshops[3] held in 2006, 2007, and 2008 that contain speakers from Spain and Latin America.

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

Table 1 Characteristics of the MAVIR database: number of word occurrences (#occ.), duration (dur.) in minutes (min), number of speakers (#spk.), and average MOS (Ave. MOS)

File ID	Data	#occ.	dur. (min)	#spk.	Ave. MOS
Mavir-02	train	13,432	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	11,179	70.05	1 (1 ma.)	2.39
Mavir-12	train	11,168	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	43,467	260.27	13 (12 ma. 1 fe.)	2.56
ALL	dev	10,512	59.96	4 (3 ma. 1 fe.)	2.64
ALL	test	20,277	121.3	6 (5 ma. 1 fe.)	2.65

These characteristics are displayed for training (train), development (dev), and testing (test) datasets

The MAVIR Spanish data consist of spontaneous speech files from different speakers, which amount to about 7 h of speech. These data are then 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[4].

The speech data were originally recorded in several audio formats (pulse-code modulation (PCM) mono and stereo, MP3, 22.05 kHz, and 48 kHz, among others). The recordings were converted to PCM, 16 kHz, single channel, and 16 bits per sample using the SoX tool[5]. All the recordings (except one) were made with the same equipment, a Digital TASCAM DAT model DA-P1. Different microphones were used, which 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 controlled at all, but in most cases the distance was smaller than 50 cm. 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

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

Table 2 Characteristics of the RTVE database: number of word occurrences (#occ.), duration (dur.) in minutes (min.), number of speakers (#spk.), and average MOS (Ave. MOS)

File ID	Data	#occ.	dur. (min)	#spk.	Ave. MOS
LN24H-20151125	dev2	21,049	123.50	22	3.37
LN24H-20151201	dev2	19,727	112.43	16	3.27
LN24H-20160112	dev2	18,617	110.40	19	3.24
LN24H-20160121	dev2	18,215	120.33	18	2.93
millennium-20170522	dev2	8330	56.50	9	3.61
millennium-20170529	dev2	8812	57.95	10	3.24
millennium-20170626	dev2	7976	55.68	14	3.55
millennium-20171009	dev2	9863	58.78	12	3.60
millennium-20171106	dev2	8498	59.57	16	3.40
millennium-20171204	dev2	9280	60.25	10	3.29
millennium-20171211	dev2	9502	59.70	12	2.95
millennium-20171218	dev2	9386	55.55	15	2.70
EC-20170513	test	3565	22.13	N/A	3.12
EC-20170520	test	3266	21.25	N/A	3.38
EC-20170527	test	2602	17.87	N/A	3.42
EC-20170603	test	3527	23.87	N/A	3.90
EC-20170610	test	3846	24.22	N/A	3.31
EC-20170617	test	3368	21.55	N/A	3.36
EC-20170624	test	3286	22.60	N/A	3.65
EC-20170701	test	2893	22.52	N/A	3.47
EC-20170708	test	3425	23.15	N/A	3.58
EC-20170715	test	3316	22.55	N/A	3.82
EC-20170722	test	3929	27.40	N/A	3.88
EC-20170729	test	4126	27.45	N/A	3.61
EC-20170909	test	3063	21.05	N/A	3.64
EC-20170916	test	3422	24.60	N/A	3.40
EC-20170923	test	3331	22.02	N/A	3.24
EC-20180113	test	2742	19.02	N/A	3.80
EC-20180120	test	3466	21.97	N/A	3.28
EC-20180127	test	3488	22.52	N/A	3.56
EC-20180203	test	3016	21.60	N/A	3.90
EC-20180210	test	3214	23.20	N/A	3.71
EC-20180217	test	3094	20.33	N/A	3.57
EC-20180224	test	3140	20.78	N/A	3.56
millennium-20170703	test	8714	55.78	N/A	1.10
millennium-20171030	test	8182	57.05	N/A	3.44
ALL	train	3,729,924	27729	N/A	3.04
ALL	dev1	545,952	3742.88	N/A	2.90
ALL	dev2	149,255	930.64	N/A	3.25
ALL	test	90,021	605.48	N/A	3.32

These characteristics are displayed for training (train), development (dev), and testing (test) datasets. Results for train and dev1 are not reported per file due to the large number of files (about 400 for train and about 60 for dev1)

Table 3 Characteristics of the COREMAH database: number of word occurrences (#occ.), duration (dur.) in minutes (min.), number of speakers (#spk.), and average MOS (Ave. MOS)

File ID	#word occ.	dur. (sec)	#spk.	Ave. MOS
49-50-rejection	343	109	2 (1 ma. 1 fe.)	1.90
49-50-compliment	470	126	2 (1 ma., 1 fe.)	2.35
49-50-apology	585	191	2 (1 ma., 1 fe.)	2.17
51-52-rejection	227	57	2 (2 fe.)	2.82
51-52-compliment	244	54	2 (2 fe.)	3.28
51-52-apology	283	59	2 (2 fe.)	4.02
53-54-rejection	183	47	2 (2 fe.)	3.26
53-54-compliment	152	44	2 (2 fe.)	2.58
53-54-apology	224	57	2 (2 fe.)	3.20
55-56-rejection	202	62	2 (1 ma., 1 fe.)	2.54
55-56-compliment	261	74	2 (1 ma., 1 fe.)	2.81
55-56-apology	337	82	2 (1 ma., 1 fe.)	2.46
57-58-rejection	509	153	2 (1 ma., 1 fe.)	2.62
57-58-compliment	328	89	2 (1 ma., 1 fe.)	1.65
57-58-apology	566	177	2 (1 ma., 1 fe.)	2.79
59-60-rejection	146	51	2 (2 fe.)	2.79
59-60-compliment	166	49	2 (2 fe.)	2.19
59-60-apology	167	41	2 (2 fe.)	3.54
61-62-rejection	286	74	2 (1 ma., 1 fe.)	2.27
61-62-compliment	192	46	2 (1 ma., 1 fe.)	2.99
61-62-apology	206	52	2 (1 ma., 1 fe.)	2.32
63-64-rejection	324	103	2 (1 ma., 1 fe.)	3.11
63-64-compliment	379	99	2 (1 ma., 1 fe.)	2.56
63-64-apology	437	128	2 (1 ma., 1 fe.)	2.62
65-66-rejection	252	60	2 (1 ma., 1 fe.)	2.91
65-66-compliment	188	47	2 (1 ma., 1 fe.)	2.46
65-66-apology	198	53	2 (1 ma., 1 fe.)	3.13
67-68-rejection	201	59	2 (2 fe.)	2.14
67-68-compliment	166	50	2 (2 fe.)	4.06
67-68-apology	218	63	2 (2 fe.)	3.12
69-70-rejection	99	33	2 (2 fe.)	4.07
69-70-compliment	89	30	2 (2 fe.)	2.43
69-70-apology	127	46	2 (2 fe.)	4.30
71-72-rejection	360	110	2 (1 ma., 1 fe.)	2.17
71-72-compliment	257	72	2 (1 ma., 1 fe.)	2.61
71-72-apology	328	93	2 (1 ma., 1 fe.)	2.06
ALL	9700	2740	24 (7 ma., 17 fe.)	2.46

These characteristics are displayed for training (train), development (dev), and testing (test) datasets

settings and the variety of phenomena in the spontaneous speech in this database make it appealing and challenging enough for the evaluation.

2.3.2 RTVE

The RTVE database belongs to the broadcast news domain and contains speech from different television (TV) programs recorded from 2015 to 2018 (e.g., Millenium, La tarde en 24H, Comando actualidad, España en comunidad, to name a few). These comprise about 570 h in total, which were further divided into training, development, and test sets for the purpose of this evaluation. To prepare the data for the evaluation, organizers manually added the timestamps for the roughly 2700 occurrences of the spoken terms used in the development and test evaluation sets. The training data were available to participants with the corresponding subtitles of the speech data (though these could contain non-accurate word transcriptions), and the development data were further divided into two different development sets, as follows: The *dev1* dataset consists of about 60 h of speech material with human-revised word transcriptions without time alignment. The *dev2* dataset, which was employed as *real* development data for STD evaluation, consists of 15 h of speech data. The recordings were provided in Advanced Audio Coding (AAC) format, stereo, 44.1 kHz, and variable bit rate. As far as we know, this database represents the largest speech database employed in any Spanish SoS evaluation. More information about the RTVE database can be found in [99].

2.3.3 COREMAH

The COREMAH database contains conversations about different topics such as rejection, compliment, and apology, which were recorded in 2014 and 2015 in a university environment[6]. This database contains Spanish recordings from speakers with different levels of Spanish (native, advanced C1, and intermediate B1). Since the main purpose of this database is to evaluate the submitted systems to an unseen domain, only the Spanish native speaker recordings are employed in the evaluation to recreate the same conditions of the other databases. The speech data amount to about 45 min. To prepare the data for the evaluation, organizers manually added the timestamps for the roughly 1000 occurrences of the spoken terms used in the test evaluation set.

The original recordings are videos in the Moving Picture Experts Group (MPEG) format. The audio of these videos was extracted and converted to PCM, 16 kHz, single channel, and 16 bits per sample using the *ffmpeg*[7] tool. It is worth mentioning that this database contains a high degree of overlapped speech, which makes it quite challenging.

[6] http://www.lllf.uam.es/coremah/[100]
[7] https://ffmpeg.org/

2.3.4 Term list selection

The selection of terms for the development and test sets aimed to build a realistic scenario for STD, by including high occurrence terms, low occurrence terms, in-language (INL) (i.e., Spanish) terms, out-of-language (OOL) (i.e., foreign) terms, single-word and multi-word terms, in-vocabulary and out-of-vocabulary terms, and terms of different length. 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 lists such as the number of INL and OOL terms, the number of single-word and multi-word terms, and the number of INV and OOV terms, along with the number of occurrences of each set in the corresponding speech database. It must be noted that a multi-word term is considered OOV in case any of the words that form the term is OOV.

2.4 Comparison to other STD international evaluations

Spoken Term Detection evaluations have been organized from more than a decade. In 2006, the NIST launched the first NIST STD evaluation [95], with English, Mandarin Chinese, and Modern Standard and Levantine Arabic as target languages. The speech 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). NIST publicly released the results of this evaluation, and they are summarized in Table 5.

A significant amount of STD research has been carried out in the framework of the IARPA BABEL program and NIST Open Keyword Search (OpenKWS) evaluation series [19, 28, 30, 32, 33, 44, 52, 56, 58, 63, 65–67, 70, 101–107]. The BABEL program was born in 2011 aiming to develop fully automatic and noise-robust speech recognition systems in a limited time (e.g., one week) and with a limited amount of transcribed training data. This program supports research in low-resource languages such as Cantonese, Pashto, Tagalog, Turkish, Vietnamese, Swahili, and Tamil, among others. From 2013 to 2016, NIST organized an annual STD evaluation called OpenKWS, which is included within the BABEL program, but open to other research groups besides BABEL participants [108–111]. This evaluation was quite similar to the former NIST STD 2006 evaluation and included 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 2017, NIST also launched the biennial Open Speech Analytics Technologies (OpenSAT) evaluation series, which includes keyword search among its tasks. This series goal is "to provide broad support for the advancement of speech analytic technologies by including multiple speech analytic tasks and multiple data domains."

Table 4 Development and test term list characteristics for MAVIR, RTVE, and COREMAH databases

Term list	dev-MAVIR	dev-RTVE	test-MAVIR	test-RTVE	test-COREMAH
#IN-LANG terms (occ.)	354 (959)	307 (1151)	208 (2071)	301 (1082)	153 (1022)
#OUT-LANG terms (occ.)	20 (55)	91 (351)	15 (50)	103 (162)	8 (16)
#SINGLE terms (occ.)	340 (984)	380 (1280)	198 (2093)	383 (1186)	145 (1004)
#MULTI terms (occ.)	34 (30)	18 (222)	25 (28)	21 (58)	16 (34)
#INV terms (occ.)	292 (668)	312 (1263)	192 (1749)	316 (1035)	128 (948)
#OOV terms (occ.)	82 (346)	86 (239)	31 (372)	88 (209)	33 (90)

"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 lists varies between 4 and 27 graphemes. The term length of the mAVIR and rTVE test term lists varies between 4 and 28 graphemes. The term length of the cOREMAH test term list varies between 3 and 17 graphemes

This evaluation focused on low-resources languages, as the previous OpenKWS, and speech data comprised conversational telephone speech.

In the ALBAYZIN 2018 STD evaluation, the audio comprises diverse recording conditions: (1) real talks in real workshops held in large conference rooms with public, (2) conversational speech, and (3) broadcast news speech. In the recordings of the workshops, microphones, conference rooms, and even recording conditions change from one recording to another, and tabletop and ground standing microphones were typically employed. In addition, our evaluation explicitly defines different in-vocabulary and out-of-vocabulary term sets. 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/OpenKWS/OpenSAT evaluations.

STD evaluations have also been held in the framework of the NTCIR conferences from 2011 to 2016 [112–115]. Data used in these evaluations are 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. In these evaluations, the organizers provided the participants with manual transcriptions of the speech data and the output of an LVCSR system. Table 7 presents the best result obtained in each evaluation, where the F-measure was used as the evaluation metric. Although MAVIR data employed in 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 a different language, employs a larger list of terms along with three different databases (each covering a different domain), and defines disjoint development and test term lists to measure the generalization capability of the systems. Besides, the evaluation metric used in these evaluations is different. All these differences make system comparison very difficult.

2.5 Comparison to previous ALBAYZIN search on speech evaluations

From 2012, ALBAYZIN STD evaluation has been integrated within the framework of ALBAYZIN SoS evaluation. This SoS evaluation includes two different tasks, named STD and QbE STD. In 2012, participants focused on the QbE STD task, whereas in 2014, systems were mainly submitted to the STD task. In 2016 and 2018, both tasks received the same attention from participants. Specifically, from 2014, the ALBAYZIN STD evaluation has evolved in different aspects:

- *Evaluation domains.* In the evaluation held in 2014, a single domain (spontaneous speech from workshop talks) was chosen. In 2016, a novel Spanish database (Spanish European parliament sessions) was added to that domain, which allowed measuring the system performance on an additional dataset, for which neither training nor development data were provided. The system ranking was based on the performance obtained on the workshop talk domain. On the other

Table 5 Best performance (in terms of Actual Term Weighted-Value, ATWV) obtained in the NIST STD 2006 evaluation for 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

Table 6 Best performance (in terms of Actual Term-Weighted Value, ATWV) obtained in the different editions (2013, 2014, 2015, and 2016) of the OpenKWS evaluations under the full language pack condition

Evaluation	ATWV	Language
OpenKWS 2013	0.6248	Vietnamese
OpenKWS 2014	0.5802	Tamil
OpenKWS 2015	0.6548	Swahili
OpenKWS 2016	0.8730	Georgian

Table 7 Best performance (in terms of F-measure) obtained in the different editions of 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

hand, the evaluation held in 2018 employs three different domains: spontaneous speech from workshop talks, broadcast news (which contains the largest database used in any Spanish STD evaluation), and spontaneous speech from 2-people conversations (for which neither training nor development data were provided). This makes the evaluation more attractive for participants, since they can evaluate the submitted systems in different domains and conditions. Moreover, since the evaluation ranking is based on the average performance from the different datasets, participants were encouraged to build multi-domain STD systems. This represents the most important difference compared to the evaluation held in 2016.

- *Number and complexity of search terms.* The number of terms used in the evaluations increases from one evaluation to another. In 2014, there were 548 terms for searching, 780 terms in 2016, and 1560 terms in the 2018 evaluation. In addition, the complexity of the search terms has also increased from one evaluation to another, with more out-of-language, multi-word, and OOV terms.
- *Evaluation sets.* In 2014, the evaluation organizers provided two different datasets: training/development and test datasets. Aiming to solve the system bias to the training data when the submitted systems were evaluated on development data, the 2016 and 2018 evaluations provided three different datasets: training, development, and test.

2.6 Participants

Three different teams submitted 11 different systems to the ALBAYZIN 2018 Spoken Term Detection evaluation, as listed in Table 8. About 3 months were given to the participants for system development and, therefore, the STD evaluation focuses on building STD systems in a limited period of time. The training, development, and test data were released to the participants at different times. Training and development data were released on June 30, 2018. The test data were released on September 24, 2018. The final system submission was due on October 21, 2018. Final results were discussed at IberSPEECH 2018 conference on November 21, 2018.

Table 8 Participants in the ALBAYZIN 2018 STD evaluation along with the systems submitted

Team ID	Research institution	Systems	Type of system
GTM-IRLab	AtlantTIC Research Center+Information Retrieval Lab. Universidade de Vigo+Universidade da Coruña, Spain	Combined Kaldi Proxy Kaldi Phone-based	LVCSR+phone-based LVCSR Phone-based
CENATAV	Voice Group, Advanced Technologies Application Center, Cuba	Kaldi-DNN Kaldi-SGMM Kaldi-GMM	LVCSR LVCSR LVCSR
GTTS	Universidad del País Vasco, Spain	Combined Synt-DTW Super-BNF Synt-DTW Multilingual-BNF Synt-DTW Monoph.-BNF Synt-DTW Triph.-BNF Synt-DTW	QbE-STD QbE-STD QbE-STD QbE-STD QbE-STD

3 Systems

In this section, the systems submitted to the evaluation are described (see Appendix). These systems can be divided into three different categories, as presented in Table 8: (1) LVCSR-based approaches, (2) subword approaches based on phone units, and (3) generating a spoken query from the written form of the term using speech synthesis and employing dynamic time warping (DTW)-based search in a QbE-STD framework.

3.1 Combined Kaldi-based STD system (Combined Kaldi)

This system combines a word-based STD system and a phone-based STD system, 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 [73] and employs deep neural network (DNN)-based acoustic models. Specifically, a DNN-based context-dependent speech recognizer is trained following the DNN training approach presented in [116]. Forty-dimensional Mel-frequency cepstral coefficients (MFCCs) augmented with three pitch- and voicing-related features [117] and appended with their delta and acceleration coefficients are first extracted for each speech frame. The DNN has 6 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, whereas 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 [118] using these DNN-based acoustic models.

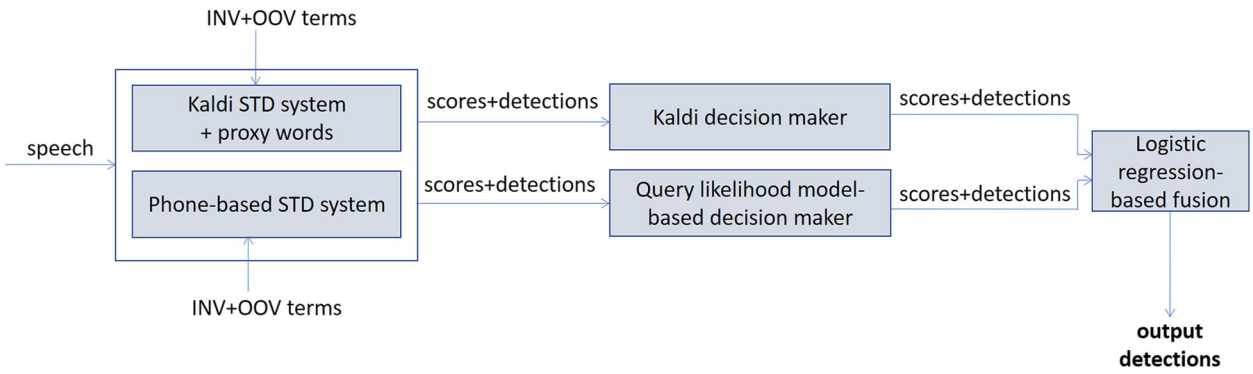

Fig. 2 Architecture of the Combined Kaldi STD system

The data used to train the acoustic models of this Kaldi-based LVCSR system are extracted from the Spanish training material of the 2006 TC-STAR automatic speech recognition evaluation campaign[8], which amounts to about 99 h of speech, and the Galician broadcast news database Transcrigal [119], which amounts to about 26 h of speech. 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 are discarded, so in the end the acoustic training material consists of approximately 104.5 h.

The language model employed in the LVCSR system is constructed using a text database of 150 million 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] [120]). Specifically, the LM is obtained from static interpolation of two 4-gram-based language models which are trained using these different text databases. Both LMs are built using the Kneser-Ney discounting strategy employing the SRILM toolkit [121], and the final LM is obtained using the SRILM static *n*-gram interpolation functionality. One of the 4-gram LMs is trained from the subtitles provided by the evaluation organizers within the RTVE training data, and the other LM is built from the rest of the text corpora. Both LMs contain 15 million 4-grams, 10 million 3-grams, 780K 2-grams, and 300K 1-grams. The LM vocabulary size is limited to the most frequent 300K words and, for each evaluation data set, the OOV terms are removed from the language model. Grapheme-to-phoneme conversion is carried out with the Cotovia software [122].

[8] http://www.tc-star.org
[9] http://cartago.lllf.uam.es/mavir/index.pl?m=descargas

The STD subsystem integrates the Kaldi term detector [73–75], 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 [123] so that the lattices of all the utterances in the search collection are converted from individual weighted finite state transducers (WFSTs) 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 3-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 [49]. To do so, the score for each detection is computed as follows:

$$p > \frac{N_{\text{conf}}}{\frac{T}{\beta} + \frac{\beta-1}{\beta} N_{\text{conf}}}, \tag{4}$$

where p is the confidence score of the detection, N_{conf} is the sum of the confidence score of all the detections of the given term, β is set to 999.9 (as in Eq. 1), and T is the length of the audio in seconds.

The proxy words strategy in the Kaldi open-source toolkit [76] is employed for OOV term detection. This strategy consists in substituting each OOV word of the search term with acoustically similar INV proxy words so that the search of OOV terms can be carried out using the obtained INV term or terms.

3.1.2 Phone-based STD system

The phone-based STD system is applied for INV and OOV term detection and follows a probabilistic retrieval

model for information retrieval. This model consists of the following stages:

- *Indexing*. The *lattice-to-phone-lattice* tool in the Kaldi [73] toolkit is employed to produce phone lattices from the word lattices output by the LVCSR system described above. Then, 40 n-best lists are created from the phone lattices and indexed in terms of phone n-grams of different size [124, 125]. The minimum and maximum sizes of the n-grams are set to 1 and 5, respectively, according to [125]. According to the probabilistic retrieval model used in this system, each spoken document is represented by means of a language model [126]. In this case, given that the phone transcriptions have errors, several hypotheses for each transcription are used to improve the quality of the language model. The start time and duration of each phone are also stored in the index.
- *Search*. A phonetic transcription of the term is first obtained using the grapheme-to-phoneme model for Spanish included in the Cotovia software [122]. Then, the term is searched within the different indices, and a score for each spoken document is computed following the query likelihood retrieval model [127]. It must be noted that this model sorts the spoken documents according to how likely it is that they contain the term, but the start and end times of the match are required in this task. To obtain these times, the phone transcription of the term T is aligned to that of the spoken document D by computing its minimum edit distance (MED) $MED(T, D)$. This allows the recovery of the start and end times, since they are stored in the index. In addition, the MED is used to penalize the score returned by the query likelihood retrieval model (Lopez-Otero et al.: Probabilistic information retrieval models for query-by-example spoken document retrieval, submitted) (i.e., $\text{score}_{\text{LM}}(T, D)$), as follows:

$$\text{score}(\text{T}, \text{D}) = \text{score}_{\text{LM}}(T, D) \cdot \text{score}_{\text{MED}}(T, D), \tag{5}$$

  where $\text{score}_{\text{MED}}(T, D)$ is a score between 0 and 1 derived from $\text{MED}(T, D)$ and computed as

$$\text{score}_{\text{MED}}(T, D) = \frac{n_{\text{T}} - \text{MED}(T, D)}{K}, \tag{6}$$

  where n_{T} is the number of phonemes of the term, and K is the length of the best alignment path.

Indexing and search are performed using Lucene.[10]

3.1.3 Fusion

Discriminative calibration and fusion [128] are applied in order to combine the outputs of the word and phone-based STD systems described above. The global minimum score produced by the system for all the terms is used to hypothesize the missing scores. After normalization, calibration and fusion parameters are estimated by logistic regression on the development dataset to obtain improved discriminative and well-calibrated scores [129]. Calibration and fusion training are performed using the Bosaris toolkit [130].

The decision threshold, weight of the LM in the word-based system, and number of n-best lists in the phone-based system for MAVIR and RTVE development data are tuned for each dataset from the individual development dataset. However, for all the test data (i.e., MAVIR, RTVE, and COREMAH), these parameters are tuned from the combined ground-truth labels of the MAVIR and RTVE development data, aiming to avoid overfitting issues. The rest of the parameters are set based on preliminary experiments.

3.2 Kaldi+proxy words-based STD system (Proxy Kaldi)

This system is the word-based STD system described in the Section 3.1.

3.3 Phone-based sTD system (Phone-based)

This system is the phone-based STD system explained in the Section 3.1.

3.4 Kaldi-based DNN system (Kaldi-DNN)

This system, whose architecture is presented in Fig. 3, is based on an LVCSR system constructed with the open-source Kaldi toolkit [73]. Specifically, the design of the system relies on the use of the s5 Wall Street Journal (WSJ) recipe in Kaldi[11]. The acoustic features used are 13 MFCCs with cepstral mean and variance normalization (CMVN) to reduce the effects of the channel. Linear discriminant analysis (LDA), maximum likelihood linear transform (MLLT), and feature-space maximum likelihood linear regression (fMLLR) were also applied to obtain more robust features. The training of the acoustic models begins with a flat initialization of context-independent phone HMMs. Then, several re-training and alignment of acoustic models are performed to obtain context-dependent phone HMMs, following the standard procedures of the Kaldi s5 WSJ recipe[12].

These phone models consist of three HMM states, each in a tied-pdf cross-word tri-phone context with Gaussian mixture models (GMMs). Then, the GMM-HMM model is speaker-adapted by means of subspace Gaussian mixture model (SGMM), as described in [131], using fMLLR features and sharing the same

[10] http://lucene.apache.org

[11] https://github.com/kaldi-asr/kaldi/tree/master/egs/wsj/s5

[12] https://github.com/kaldi-asr/kaldi/tree/master/egs/wsj/s5

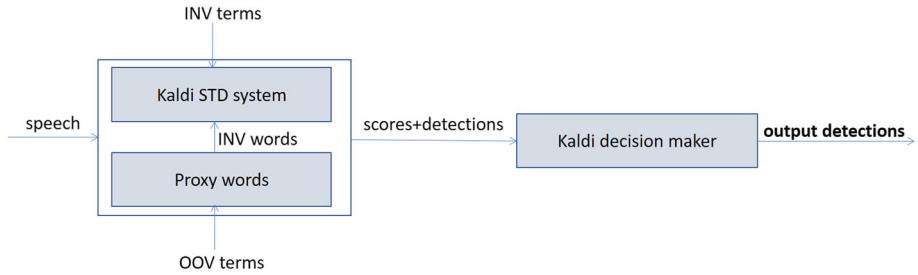

Fig. 3 Architecture of the Kaldi-DNN STD system

Gaussian model. The GMM-HMM also produces the alignments for training the DNN-based acoustic model (DNN-HMM). The DNNs contain 2 hidden layers with 300 nodes each. The number of spliced frames is 9 to produce 360 dimensional vectors as input to the first layer. The output layer is a soft-max layer representing the log-posteriors of the context-dependent states.

The data used to train the acoustic models comprise the TC-STAR data recorded from 2005 to 2007, which contain more than 26 h of speech; a subset of the *dev1* set of the RTVE data, which amounts to about 14 h of speech; and the MAVIR training data, which amount to more than 4 h of speech. In total, there are 45 h of speech material in the three datasets. Overlapped speech is removed from the *dev1* set of RTVE, so eventually 44 h of speech are used for acoustic model training.

The data used for language model training include the text transcriptions of the data used for acoustic model training, which contain 425K word occurrences. Specifically, these text transcriptions are given to the SRILM toolkit [121] to create a trigram-based LM, which consists of 38K trigrams, 155K bigrams, and 23K unigrams. The system vocabulary consists of the different words corresponding to the training data which, after removing the OOV words, amounts to 23K words. The multilingual G2P transcriber[13] is employed to obtain the phone transcription of each word.

The Kaldi decoder generates word lattices using the DNN-HMM based acoustic models. The STD subsystem, which takes the word lattices as input, includes the Kaldi term detector and Kaldi decision maker-explained in the *Combined Kaldi* system.

The proxy words strategy in the Kaldi open-source toolkit [76] is employed for OOV term detection.

All the system parameters are selected based on preliminary experiments, and no additional tuning from development data is carried out.

[13] https://github.com/jcsilva/multilingual-g2p

3.5 Kaldi-based SGMM system (Kaldi-SGMM)

This system is the same as the *Kaldi-DNN* system but SGMMs are employed for acoustic modeling in the Kaldi-based LVCSR system.

3.6 Kaldi-based GMM system (Kaldi-GMM)

This system is the same as the *Kaldi-DNN* system but GMMs are employed for acoustic modeling in the Kaldi-based LVCSR system.

3.7 Combined synthetic-Speech DTW system (Combined synt-DTW)

This system, whose architecture is shown in Fig. 4, aims to completely overcome the OOV word issue of text-based approaches. To do so, the written form of the term is synthesized to generate a spoken query that is then given to a QbE-STD system to hypothesize detections.

3.7.1 Generation of multiple spoken queries

Two different text-to-speech (TTS) tools are used for spoken query generation: the Google TTS (gTTS) Python library and command-line interface (CLI) tool [132], which provides two different female voices (es-ES and es-US); and the Cocoa TTS interface in MacOS [133], which has five different voices (three male, two female) including both European and American Spanish. In this way, for each textual form of the term, seven spoken queries $q_1, q_2, \ldots, q_7$ are synthesized.

3.7.2 Voice activity detection (VAD)

The synthesized spoken queries and the audio documents are given to a VAD system. Specifically, the Python interface for the VAD module developed by Google for the WebRTC project [134] is employed. This VAD strategy is based on Gaussian distributions of speech and non-speech features.

3.7.3 Feature extraction

The feature extraction consists in stacked bottleneck feature (SBNF) computation following the BUT/Phonexia approach [135], both for the synthesized spoken queries

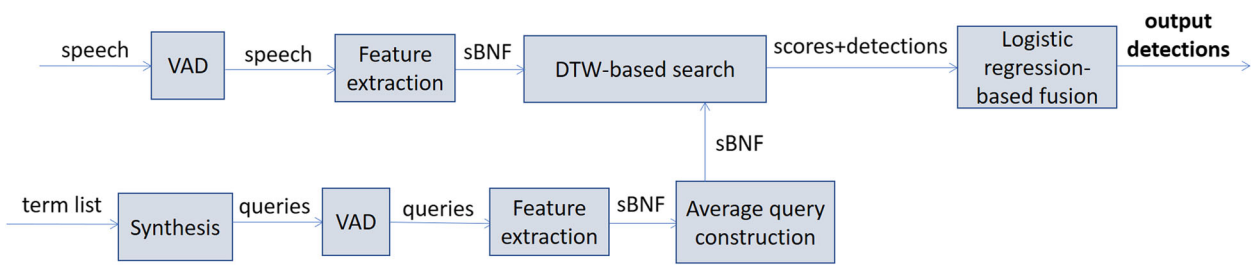

Fig. 4 Architecture of the Combined Synt-DTW STD system

and the audio documents. To do so, three different neural networks are applied, each trained to classify a different set of acoustic units and later optimized for language recognition tasks. The first network is trained on telephone speech from the English Fisher corpus [136] with 120 monophone state targets, which will be referred as FisherMono. The second one is also trained on the Fisher corpus but with 2423 triphone tied-state targets and will be referred as FisherTri. The third network is trained on telephone speech from 17 languages included in the IARPA Babel program [137], with 3096 stacked monophone state targets (BabelMulti for short). Given that the SBNF extractors are trained using 8 kHz speech signals, the documents and the synthesized spoken queries are downsampled to 8 kHz.

The architecture of the SBNF networks consists of two stages. The first stage is a standard bottleneck network fed with low-level acoustic features, which span 10 frames (100 ms), producing a bottleneck feature vector of 80 dimensions. The second stage employs five equally-spaced bottleneck feature vectors from the first stage as input and is trained on the same targets as the first stage, producing bottleneck features of the same size (80). The bottleneck features extracted from the second stage are known as stacked bottleneck features, and comprise the output of the feature extraction module. Alternatively, the extractor can output target posteriors, instead of SBNFs.

The operation of BUT/Phonexia SBNF extractors requires an external VAD module (as WebRTC VAD in our case) providing speech/non-speech information. If no external VAD is provided, a simple energy-based VAD is computed internally. This system employs the WebRTC VAD module.

The first aim for the feature extraction stage was to employ the BUT/Phonexia posteriors, but the huge size of FisherTri (2423) and BabelMulti (3096) targets requires some kind of selection, clustering or dimensionality reduction approach. Therefore, given that—at least theoretically—the same information is conveyed by sBNFs with a suitably low dimensionality (as 80 in this case), sBNFs are employed. However, this may require to pay a high price. Posteriors have a clear meaning, they can be linearly combined and their values suitably fall within the range [0,1], which makes the $-\log\cos(\alpha)$ distance also range in [0,1], where α is the angle between two vectors of posteriors. On the other hand, bottleneck layer activations have no clear meaning, it is not actually known if

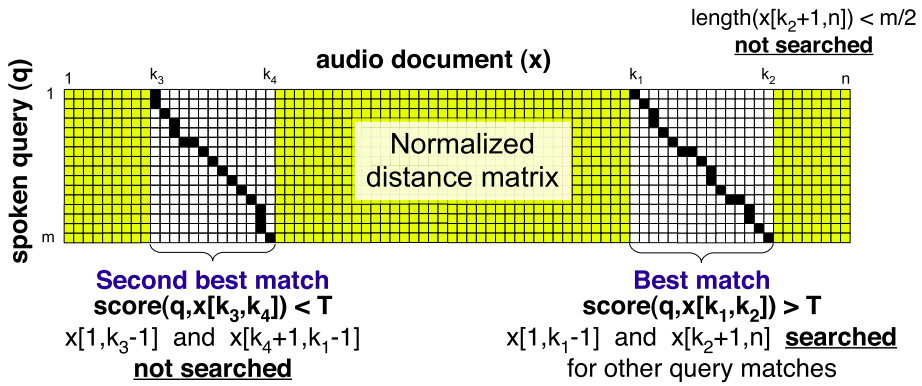

Fig. 5 Example of the iterative DTW procedure: (1) The best match of q in $x[1,n]$ is located in $x[k_1,k_2]$. (2) Since the score is greater than the established threshold T, the search continues in the surrounding segments $x[1,k_1-1]$ and $x[k_2+1,n]$. (3) $x[k_2+1,n]$ is not searched, because it is too short. (4) The best match of q in $x[1,k_1-1]$ is located in $x[k_3,k_4]$. (5) But its score is lower than T, so the surrounding segments $x[1,k_3-1]$ and $x[k_4+1,k_1-1]$ are not searched. The search procedure outputs the segments $x[k_1,k_2]$ and $x[k_3,k_4]$

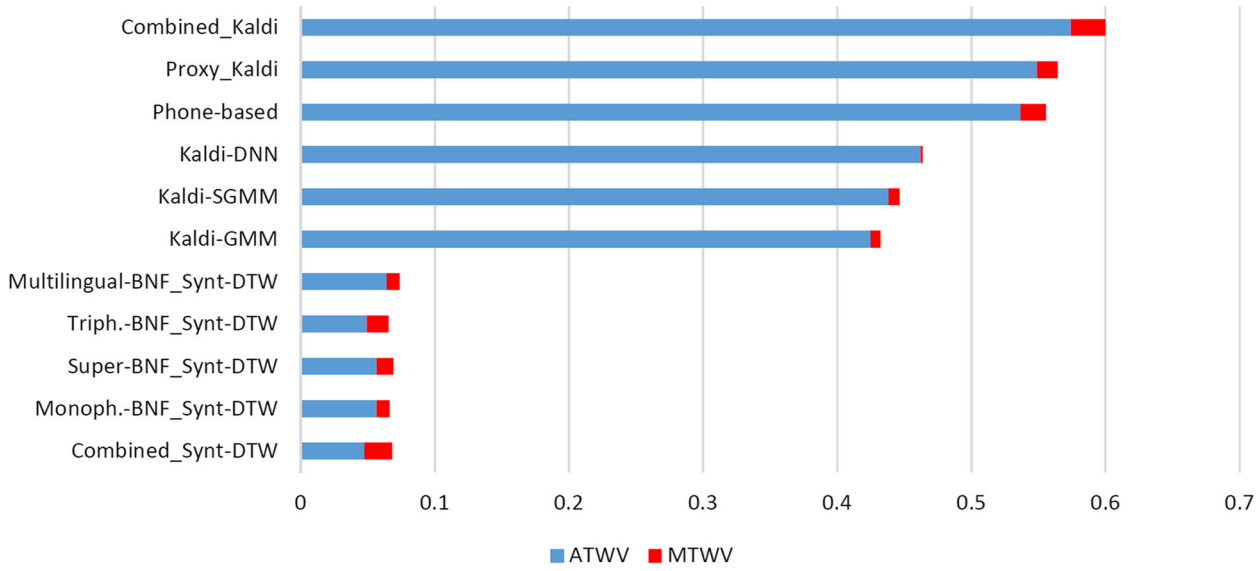

Fig. 6 Overall system results of the ALBAYZIN 2018 STD evaluation on development data

they can be linearly combined (e.g., for computing an average query from multiple query instances), and their values are unbounded, so the $-\log\cos(\alpha)$ distance no longer applies.

3.7.4 Average query construction

From the synthesized spoken queries, the longest query is taken as reference and then, optimally aligned to the other queries by means of a standard DTW procedure. Let q_l be a sequence of m_l VAD-filtered sBNF vectors for the reference query, and let q_i be the sequence of m_i vectors corresponding to another synthesized query. The alignment starts at $[1, 1]$, ends at $[m_l, m_i]$, and involves L alignments, such that each feature vector of q_l is aligned to a sequence of vectors of q_i. This is repeated for all the synthesized queries, and a set of feature vectors namely S_j is obtained from the alignment with each feature vector $q_l[j], j = 1, 2, \ldots, m_l$. Then, each $q_l[j]$ is averaged with the feature vectors in S_j to get a *single average query*, as follows:

$$q_{\text{avg}}[j] = \frac{1}{1 + |S_j|}\left(q_l[j] + \sum_{v \in S_j} v\right) \quad j = 1, 2, \ldots, m_l. \tag{7}$$

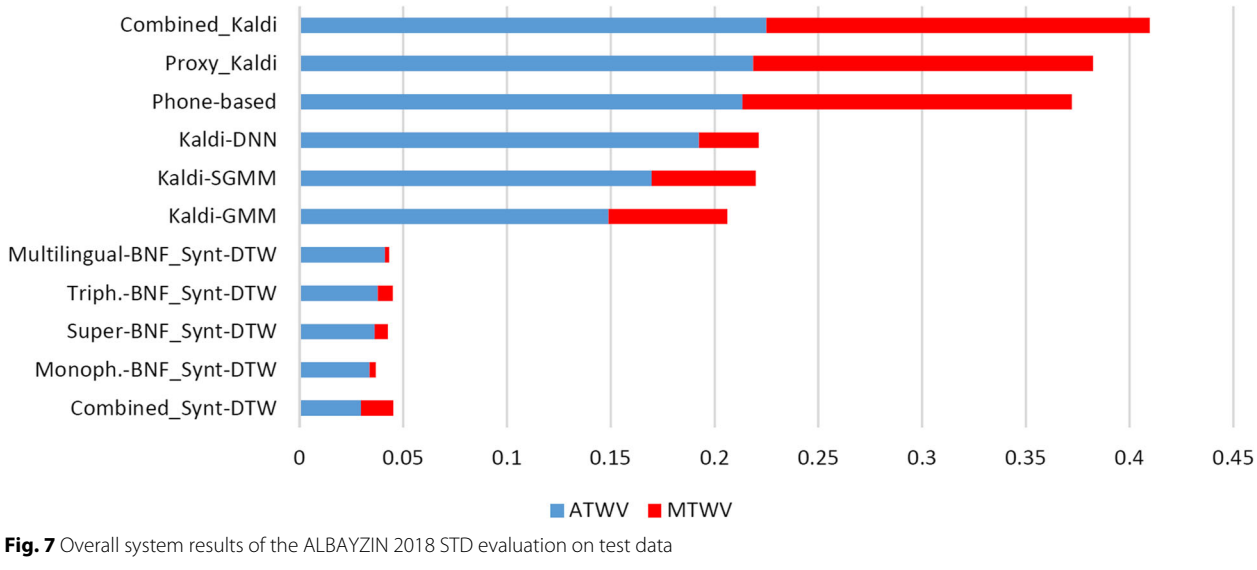

Fig. 7 Overall system results of the ALBAYZIN 2018 STD evaluation on test data

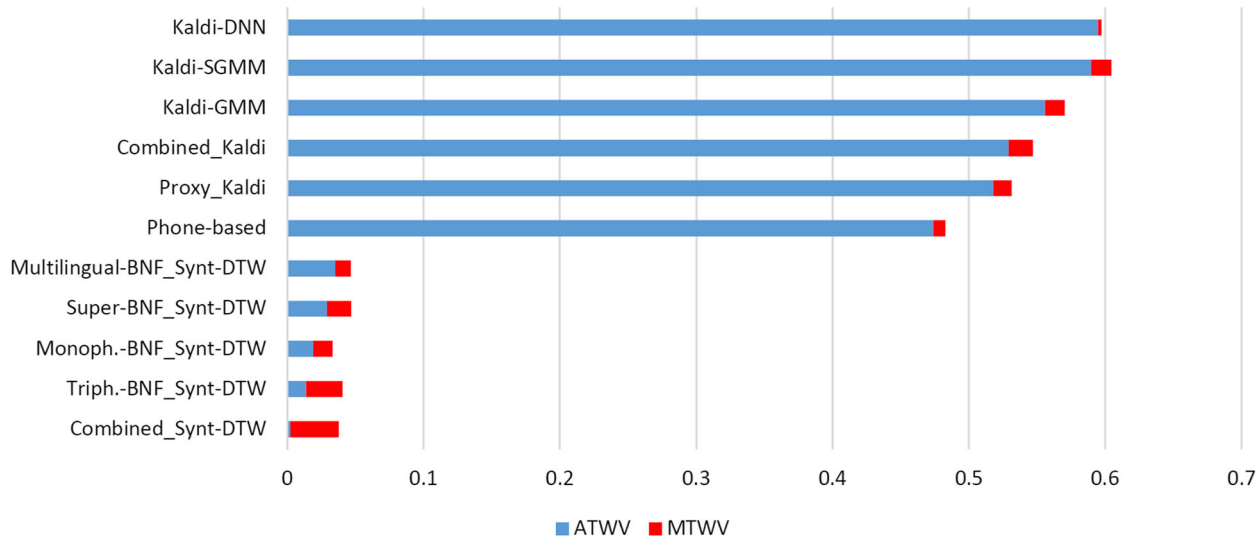

Fig. 8 System results of the ALBAYZIN 2018 STD evaluation on MAVIR development data

Finally, the average query $q_{\text{avg}}[j]$ is used to search for occurrences in the audio documents using the DTW-based approach explained next.

3.7.5 Dynamic time warping-based search

To perform the search of spoken queries in audio documents, the system follows the DTW-based approach presented in [138]. Given two sequences of sBNFs corresponding to a spoken query and an audio document, a VAD system is used to discard non-speech frames, but keeping the timestamp of each frame. To avoid memory issues, audio documents are split into chunks of 5 min with 5-s overlap and processed independently. This chunking process is key to the speed and feasibility of the search procedure.

Let $q = (q[1], q[2], \ldots, q[m])$ be the VAD-filtered sequences corresponding to a query of length m and $x = (x[1], x[2], \ldots, x[n])$ be those of an audio document of length n. Since sBNFs (theoretically) range from $-\infty$ to $+\infty$, the distance between any pair of vectors, $q[i]$ and $x[j]$, is defined as follows:

$$d(q[i], x[j]) = -\log\left(1 + \frac{q[i] \cdot x[j]}{|q[i]| \cdot |x[j]|}\right) + \log 2. \quad (8)$$

Table 9 Percentage of MAVIR INV terms that do not appear in the LVCSR system vocabulary (only for word-based STD systems)

System ID	Development OOV rate	Test OOV rate
Combined Kaldi	0.3%	5.2%
Proxy Kaldi	0.3%	5.2%
Kaldi-DNN	5.5%	20.3%
Kaldi-SGMM	5.5%	20.3%
Kaldi-GMM	5.5%	20.3%

Note that $d(v, w) \geq 0$, with $d(v, w) = 0$ if and only if v and w are aligned and pointing in the same direction, and $d(v, w) = +\infty$ if and only if v and w are aligned and pointing in opposite directions.

The distance matrix computed according to Eq. 8 is normalized with respect to the audio document x, as follows:

$$d_{\text{norm}}(q[i], x[j]) = \frac{d(q[i], x[j]) - d_{\min}(i)}{d_{\max}(i) - d_{\min}(i)}, \quad (9)$$

where:

$$d_{\min}(i) = \min_{j=1,\ldots,n} d(q[i], x[j]) \quad (10)$$

$$d_{\max}(i) = \max_{j=1,\ldots,n} d(q[i], x[j]). \quad (11)$$

In this way, matrix values are in the range $[0, 1]$ and a perfect match would produce a quasi-diagonal sequence of zeroes. This can be seen as *test normalization* since, given a query q, distance matrices take values in the same range (and with the same *relative meaning*), no matter the acoustic conditions, the speaker or other factors of the audio document x.

It must be noted that the chunking process described above makes the normalization procedure differ from that applied in [138], since $d_{\min}(i)$ and $d_{\max}(i)$ are not computed for the whole audio document but for each chunk independently. On the other hand, considering chunks of 5 min might be beneficial, since normalization is performed in a more local fashion, that is, more suited to the speaker(s) and acoustic conditions of each particular chunk.

The best match of a query q of length m in an audio document x of length n is defined as that which minimizes the average distance in a *crossing path* of the matrix d_{norm}. A crossing path starts at any given frame of x, $k_1 \in [1, n]$,

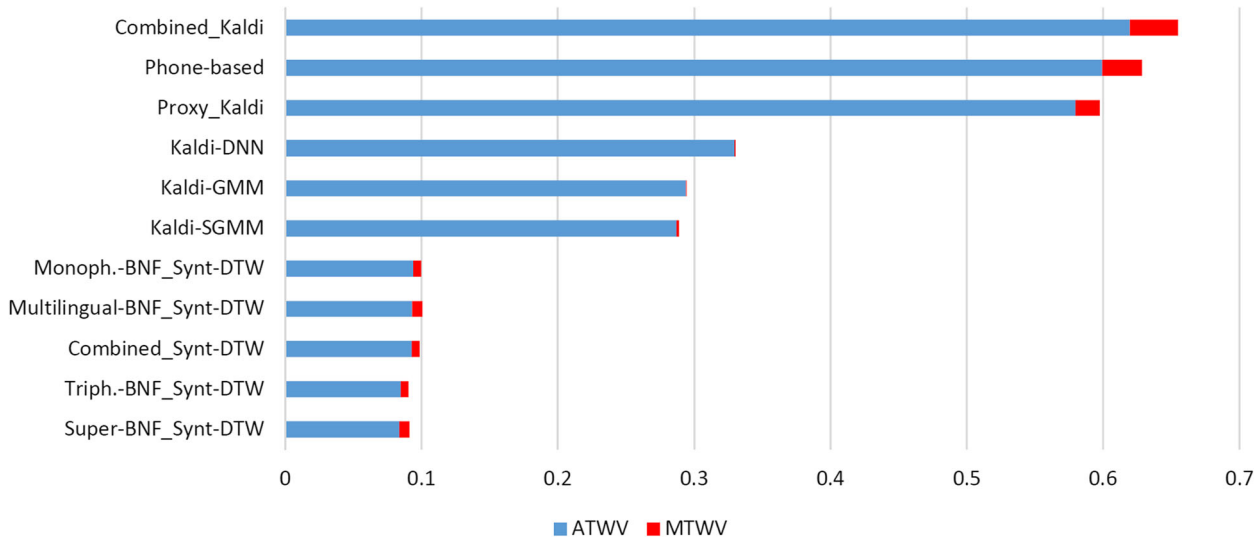

Fig. 9 System results of the ALBAYZIN 2018 STD evaluation on RTVE development data

then traverses a region of x which is optimally aligned to q (involving L vector alignments), and ends at frame $k_2 \in [k_1, n]$. The average distance in this crossing path is as follows:

$$d_{\text{avg}}(q,x) = \frac{1}{L}\sum_{l=1}^{L} d_{\text{norm}}(q[i_l], x[j_l]), \tag{12}$$

where i_l and j_l are the indices of the vectors of q and x in the alignment l, for $l = 1, 2, \ldots, L$. Note that $i_1 = 1$, $i_L = m$, $j_1 = k_1$, and $j_L = k_2$. The optimization procedure is $O(n \cdot m \cdot d)$ in time (d size of feature vectors) and $O(n \cdot m)$ in space. Readers are referred to [138] for more details.

The detection score is computed as $1 - d_{avg}(q,x)$, thus ranging from 0 to 1, being 1 only for a perfect match. The starting time and the duration of each detection are obtained by retrieving the time offsets corresponding to frames k_1 and k_2 in the VAD-filtered audio document.

This procedure is iteratively applied to find not only the best match, but also less likely matches in the same audio document. To that end, a queue of search intervals is defined and initialized with $[1, n]$. Given an interval $[a, b]$, and assuming that the best match is found at $[a', b']$, the intervals $[a, a' - 1]$ and $[b' + 1, b]$ are added to the queue (for further processing) only if the following conditions are satisfied: (1) The score of the current match is greater than a given threshold T ($T = 0.85$); (2) The interval is long enough (half the query length, $m/2$); (3) The number of matches (those already found + those waiting in the queue) is limited to less than a given threshold M ($M = 7$). An example is shown in Fig. 5. Finally, the list of matches for each query is ranked according to the scores and truncated to the N highest scores ($N = 1000$, though it effectively applied only in a few cases).

Four different DTW-based searches are carried out. Three of them employ the three sBNF sets computed in the feature extraction module (FisherMono, FisherTri, BabelMulti). The other DTW search employs the concatenation of all the three sBNF sets (which leads to 240-dimensional sBNFs). Each DTW search produces different term detections that are next fused in the fusion stage.

Table 10 Percentage of RTVE INV terms that do not appear in the LVCSR system vocabulary (only for word-based STD systems)

System ID	Development OOV rate	Test OOV rate
Combined Kaldi	5.1%	6.6%
Proxy Kaldi	5.1%	6.6%
Kaldi-DNN	52.6%	66.8%
Kaldi-SGMM	52.6%	66.8%
Kaldi-GMM	52.6%	66.8%

3.7.6 Calibration and fusion

The scores produced by the different searches are transformed according to a discriminative calibration/fusion approach commonly applied in speaker and language recognition [139].

First, the so-called *q-norm* (query normalization) is applied, so that zero-mean and unit-variance scores are obtained per query. Then, if n different systems are fused, detections are aligned so that only those supported by k or more systems ($1 \leq k \leq n$) are retained for further processing ($k = 2$). To build the full set of trials (potential detections), a rate of 1 trial per second is chosen (which is consistent with the evaluation script provided by the organizers). Given one of those detections of a query q

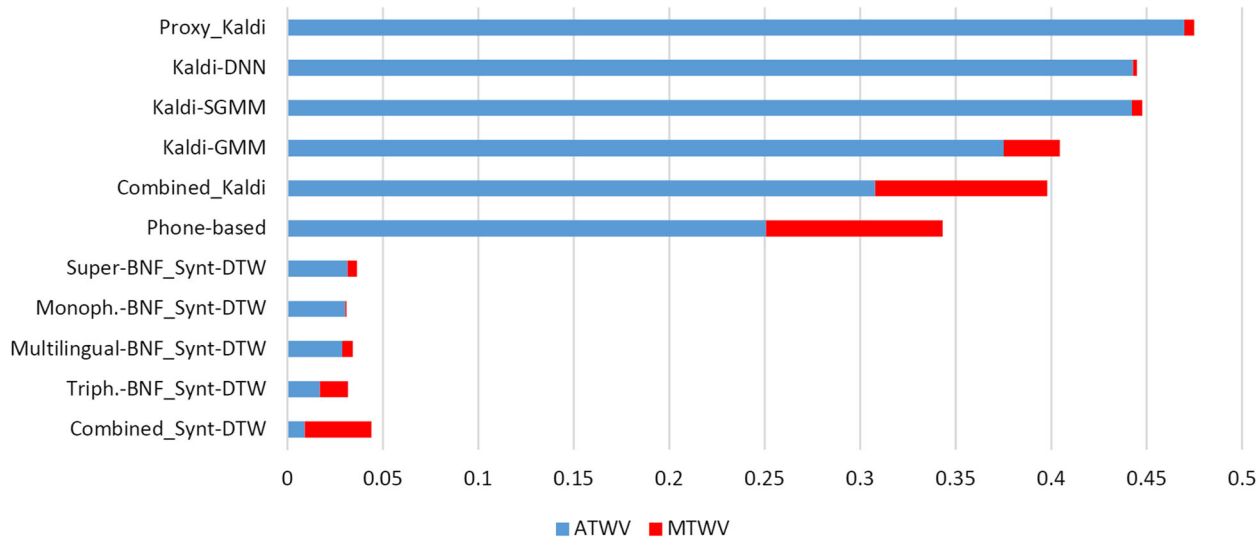

Fig. 10 System results of the ALBAYZIN 2018 STD evaluation on MAVIR test data

supported by at least k systems and a system A that did not provide a score for it, there could be different ways to fill up this *hole*. The minimum score that A has output for query q in other trials is selected. In fact, the minimum score for the query q is hypothesized for all target and non-target trials of query q for which system A has not output a detection score. When a single system is considered ($n = 1$), the majority voting scheme and the filling up of missing scores are skipped. In this way, a complete set of scores is prepared, which besides the ground truth (target/non-target labels) for a development set of queries, can be used to discriminatively estimate a linear transformation that will hopefully produce well-calibrated scores.

The calibration/fusion model is learned on the development set and then applied to both the development and test sets, using the Bosaris toolkit [130]. Under this approach, and given the effective prior (in this evaluation, $\hat{P}_{\text{target}} = C_{\text{miss}}P_{\text{target}}/(C_{\text{miss}}P_{\text{target}} + C_{\text{fa}}(1 - P_{\text{target}})) = 0.001$), the Bayes optimal threshold is applied and—at least theoretically—no further tuning would be necessary. In practice, however, if a system yields a small amount of detections, the system will be using hypothesized scores for most of the trials. As a result, the calibration/fusion

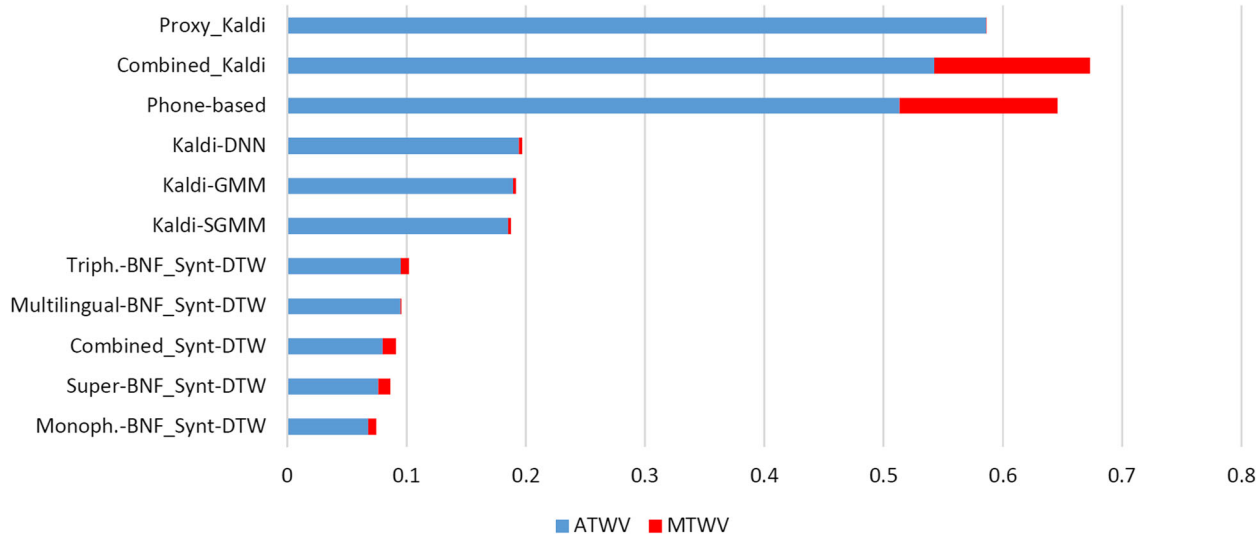

Fig. 11 System results of the ALBAYZIN 2018 STD evaluation on RTVE test data

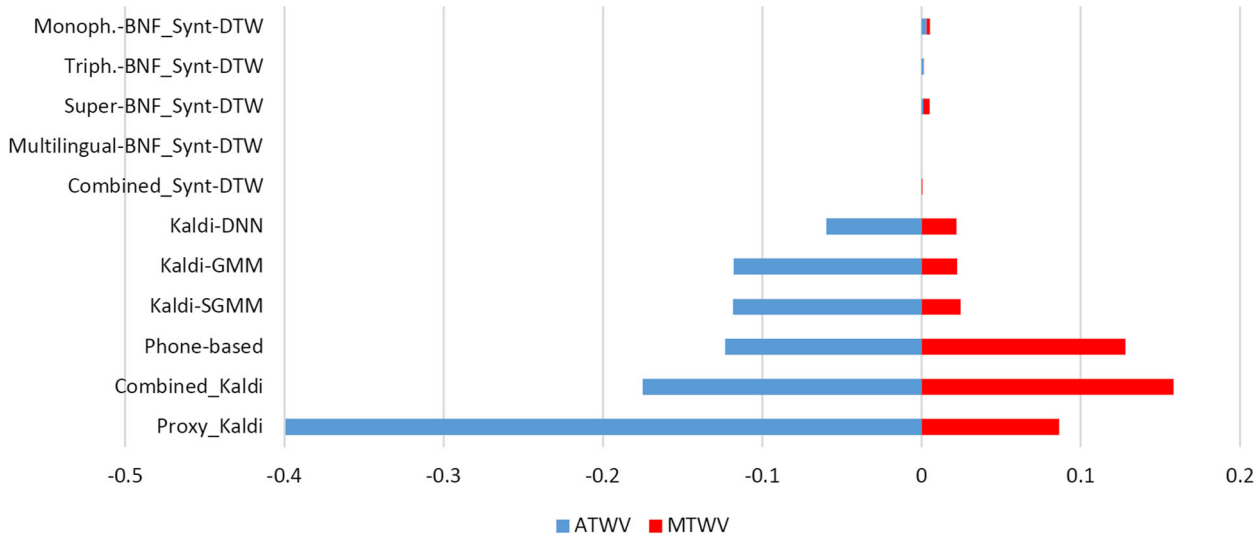

Fig. 12 System results of the ALBAYZIN 2018 STD evaluation on COREMAH test data

model would be poorly learned and the Bayes optimal threshold would not produce good results.

The calibration/fusion parameters and optimal decision threshold are obtained from the corresponding development set for each database (MAVIR and *dev2* for RTVE). For the COREMAH database, the optimal calibration/fusion parameters tuned on MAVIR data are employed, since evaluation organizers did not provide any development data for that database, and the optimal decision threshold is chosen so that 15% of the detections with the highest scores are assigned YES decision. The parameters involved in the feature extraction and search procedures are set based on preliminary experiments.

3.8 Super-bottleneck feature-based synthetic-speech DTW system (Super-BNF synt-DTW)

This system is the same as the *Combined Synt-DTW* system, except that a single DTW-based search with the concatenation of the three SBNF as features is used to hypothesize term detections.

Table 11 Percentage of COREMAH 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%
Proxy Kaldi	0%
Kaldi-DNN	14.8%
Kaldi-SGMM	14.8%
Kaldi-GMM	14.8%

3.9 Multilingual bottleneck feature-based synthetic-speech DTW system (Multilingual-BNF synt-DTW)

This system is the same as the *Super-BNF Synt-DTW* system, except that DTW-based search on the BabelMulti sBNF set is used for term detection.

3.10 Monophone bottleneck feature-based synthetic-speech DTW system (Monoph.-BNF synt-DTW)

This system is the same as the *Super-BNF Synt-DTW* system, except that DTW-based search on the FisherMono sBNF set is used for term detection.

3.11 Triphone bottleneck feature-based synthetic-speech DTW system (Triph.-BNF synt-DTW)

This system is the same as the *Super-BNF Synt-DTW* system, except that DTW-based search on the FisherTri sBNF set is used for term detection.

4 Evaluation results and discussion

4.1 Overall results

The overall evaluation results are presented in Figs. 6 and 7 for development and test data, respectively. These show that the best performance for MTWV and ATWV metrics corresponds to the *Combined Kaldi* system. Development and test data largely present the same ranking results. However, as explained next, this does not mean that the best system on development data corresponds to the best system on test data for all the databases. Different calibration threshold issues have caused this in the overall results.

4.2 Development data

4.2.1 MAVIR

System results for the MAVIR development data are presented in Fig. 8. The best performance is obtained with the *Kaldi-DNN* system, for which the small performance gap between MTWV and ATWV suggests that the threshold has been well-calibrated. This best performance is statistically significant for a paired t test ($p < 0.01$) with respect to the *Phone-based* system and the systems that employ the QbE-STD approach (i.e., *Combined Synt-DTW*, *Super-BNF Synt-DTW*, *Multilingual-BNF Synt-DTW*, *Monoph.-BNF Synt-DTW*, and *Triph.-BNF Synt-DTW*), and weakly significant ($p < 0.03$) with respect to the *Combined Kaldi* system. On the one hand, by inspecting the systems that employ a text-based STD approach, the *Phone-based* system degrades the STD performance compared with the other text-based STD systems. Although this system is based on word ASR to produce word lattices, these are then converted to phone n-grams for search, so that the word information is lost. This can be critical for highly-spontaneous and low-quality speech in MAVIR data. Nevertheless, phone-based systems typically convey fast search and indexing, and the possibility of detecting OOV terms with no additional system development. All the text-based STD systems that employ word ASR and word lattices for search do not present statistically significant differences for a paired t test, and hence they should be considered *equivalent* from an STD perspective. This indicates that the small difference in OOV rate in the development data according to Table 9 (5.5% − 0.3% = 5.2%) is not statistically significant. The systems that employ a QbE-STD approach for STD obtained a remarkably low performance. This may be due to these factors: (1) An acoustic mismatch between the synthesized queries and the test audios might lead to low scores and block the iterative DTW detection procedure. (2) The use of bottleneck layer activations as frame-level acoustic representation might be incompatible with the query averaging procedure (which worked fine with phone posteriors). (3) The absence of lexical information since no ASR system is employed.

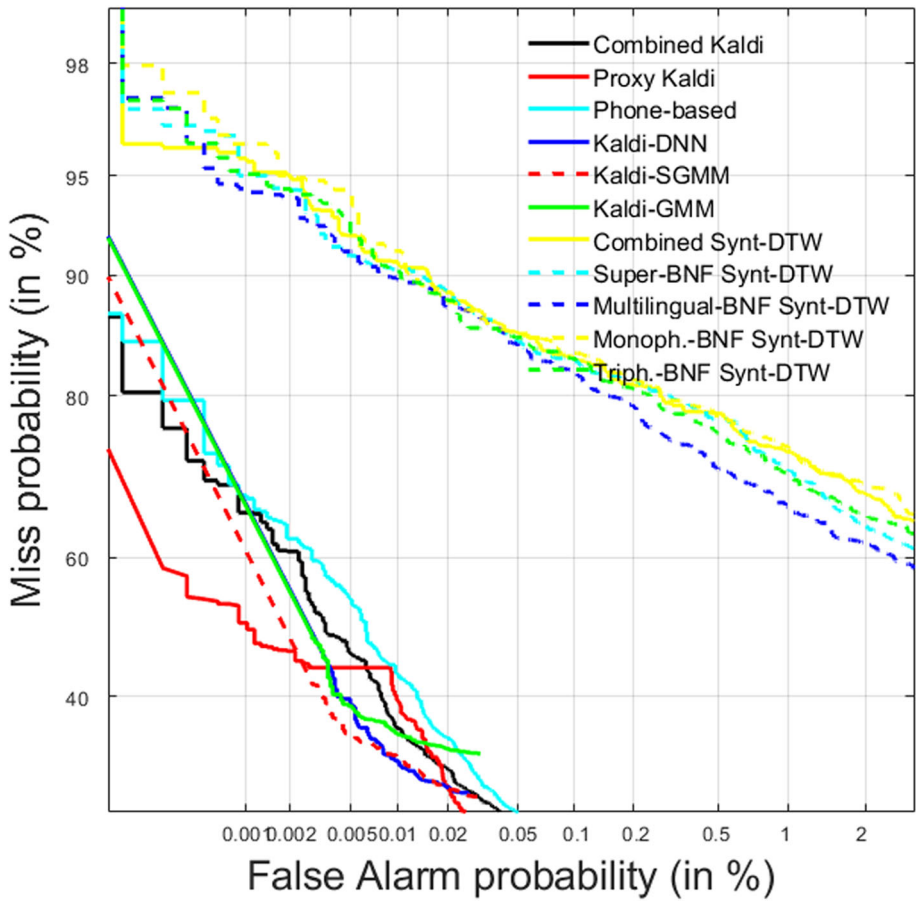

Fig. 13 The DET curves of the STD systems for MAVIR development data

4.2.2 RTVE

System results for the RTVE development data are presented in Fig. 9. The best performance is obtained with the *Combined Kaldi* system, for which the very small performance gap between MTWV and ATWV suggests that the threshold has been optimally calibrated. This best performance is statistically significant for a paired t test ($p < 0.01$) compared with the rest of the systems, except with the *Phone-based* system, for which the improvement is weakly significant ($p < 0.02$), and the *Proxy Kaldi* system, for which the performance gap is insignificant. This different ranking, compared with the MAVIR development data, is due to the large difference in the OOV rate ($52.6\% - 5.1\% = 47.5\%$) between the *Kaldi-DNN* and *Combined Kaldi* systems, as presented in Table 10. It is worth mentioning that there is no significant difference for a paired t test between the *Proxy Kaldi* and the *Phone-based* systems. This suggests that phone-based systems are able to perform similarly to word-based systems for high-quality and well-pronounced speech such as that of the RTVE data. The systems that employ the QbE-STD approach obtain the worst results, probably due to the same causes mentioned in the previous section.

4.3 Test data

4.3.1 MAVIR

System results for the MAVIR test data are presented in Fig. 10. The best performance is obtained with the *Proxy Kaldi* system, for which the performance gap between MTWV and ATWV metrics suggests that the threshold calibration works well. This best performance is statistically significant for a paired t test ($p < 0.01$) compared with all the systems except the *Kaldi-DNN* and *Kaldi-SGMM* systems. On the one hand, the low performance of the *Combined Kaldi* system indicates some calibration issues in the fusion stage. This is confirmed by the low performance obtained in the *Phone-based* system, which indicates that the parameter tuning on MAVIR development data does not generalize well in unseen data. On the other hand, the *Proxy Kaldi* system incorporates INV and OOV term detection in a common framework, and hence it is more robust against calibration issues. The differences in OOV rate shown in Table 9 between the *Proxy*

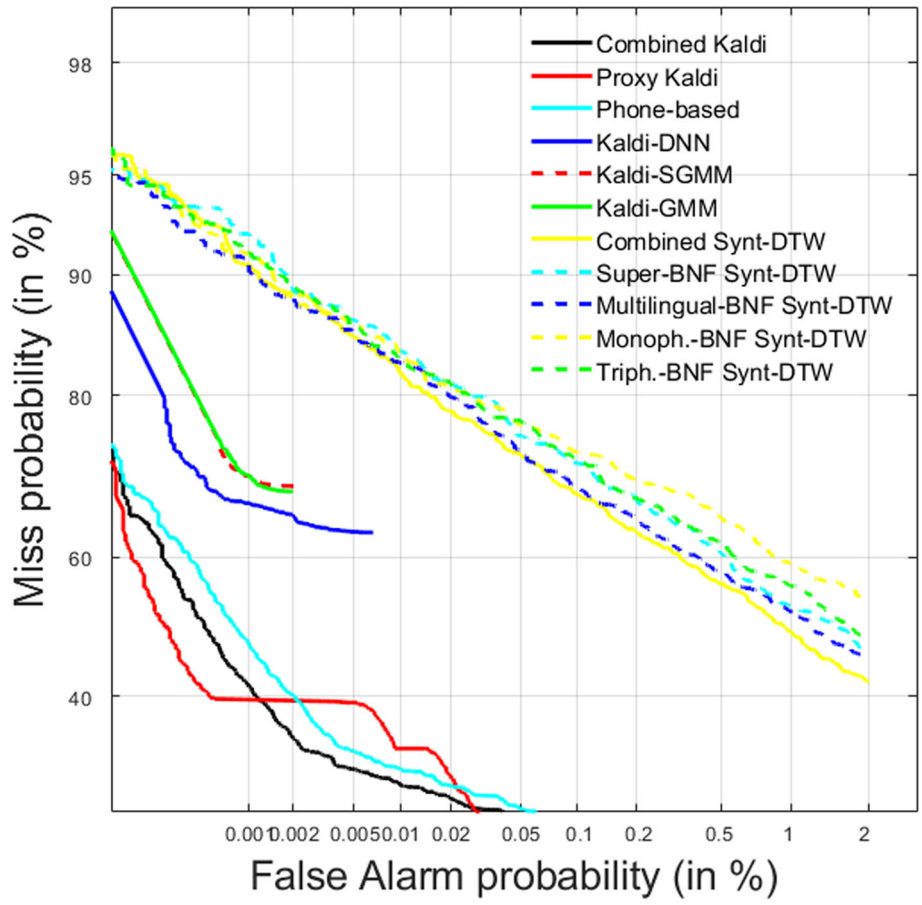

Fig. 14 The DET curves of the STD systems for RTVE development data

Kaldi, *Kaldi-DNN*, and *Kaldi-SGMM* systems (20.3% − 5.2% = 15.1%) do not produce a statistically significant reduction in the ATWV performance, which suggests that robust acoustic models along with an effective OOV term detection can mitigate the OOV issue in low-quality and highly-spontaneous speech domains. Again, the systems based on a QbE-STD approach obtained a much lower performance.

4.3.2 RTVE

System results for the RTVE test data are presented in Fig. 11. The best performance corresponds to the *Proxy Kaldi* system, for which MTWV and ATWV are very close, indicating an almost perfect threshold calibration. This best performance is statistically significant for a paired t test ($p < 0.01$) with respect to the rest of the systems, except the *Combined Kaldi* and the *Phone-based* systems. Similar findings to those of the RTVE development data arise: (1) The large difference in OOV rate shown in Table 10 for test data (66.8% − 6.6% = 62.2%) produces large differences in terms of ATWV for the word-based STD systems. (2) The results of the *Phone-based* system can be considered statistically *equivalent* to those obtained with the *Proxy Kaldi* system, which highlights the performance of the *n*-grams when facing term detection in an open-vocabulary STD system.

4.3.3 COREMAH

System results for the COREMAH test data are presented in Fig. 12. The best performance is for the *Monoph.-BNF Synt-DTW* system, although all of the systems obtained very low MTWV/ATWV results. This best performance is statistically significant for a paired t test ($p < 0.01$) compared with the *Combined Kaldi*, *Kaldi-SGMM*, and *Kaldi-GMM* systems, and weakly significant compared with the *Proxy Kaldi* ($p < 0.04$), *Phone-based* ($p < 0.02$), and *Kaldi-DNN* ($p < 0.03$) systems. The low performance obtained in these data may be due to the following factors: (1) These data contain overlapped speech, which significantly reduces ASR performance; (2) the absence of training/development data belonging to this domain, which prevents the systems from being properly tuned to these data. This is especially critical in systems based on word

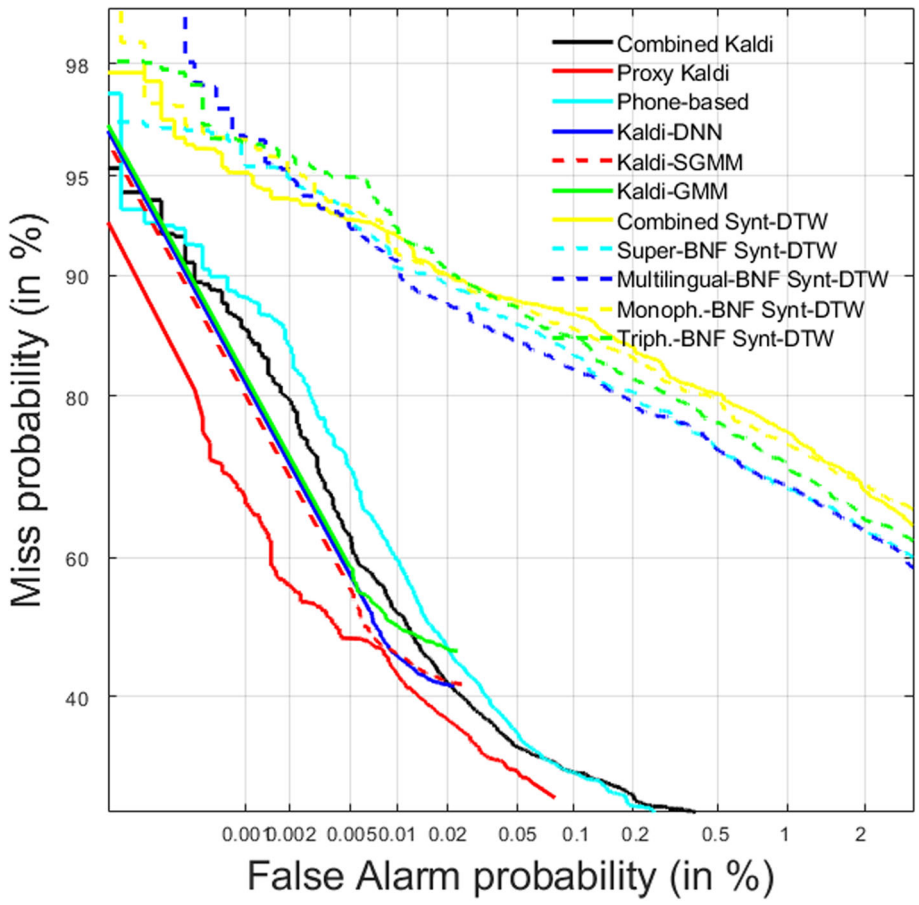

Fig. 15 The DET curves of the STD systems for MAVIR test data

speech recognition, which typically need a larger dataset for system construction than systems based on QbE-STD since these just rely on template-matching of features. For systems based on word ASR, the threshold calibration issue is more important, as can be seen from the performance gap between MTWV and ATWV. Table 11 shows the OOV rate of the word-based systems. In this case, the OOV rate is not as critical to STD performance as the change in the data domain.

4.4 Development and test data DET curves

DET curves of the systems submitted to the evaluation are presented in Figs. 13 and 14 for MAVIR and RTVE development data, respectively, and in Figs. 15, 16, and 17 for MAVIR, RTVE, and COREMAH test data, respectively.

On MAVIR development data, the *Proxy Kaldi* system performs the best for low and high FA rates, and the *Kaldi-DNN* and *Kaldi-SGMM* systems perform the best for moderate FA rates. On RTVE development data, the *Proxy Kaldi* system performs the best for low and high FA rates, and the *Combined Kaldi* system performs the best for moderate FA rates. According to the MTWV/ATWV results (see Figs. 8 and 9), this means that the best operating point is placed in moderate FA rates for both datasets.

On MAVIR test data, the *Proxy Kaldi* system performs the best for all the operation points, as expected from the MTWV/ATWV results (see Fig. 10). On RTVE test data, the *Proxy Kaldi* system performs the best for low FA rates, and the *Combined Kaldi* system performs the best for low miss rates. According to the MTWV/ATWV results (see Fig. 11), this means that the best operating point resides in low FA rates. On COREMAH test data, the *Combined Kaldi* system performs the best for low FA rates, and the *Proxy Kaldi* system performs the best for low miss rates. According to the MTWV/ATWV results (see Fig. 12), this differs from the best ATWV (which is obtained with the *Monoph.-BNF Synt-DTW* system). However, this *Monoph.-BNF Synt-DTW* system only outputs one detection as hit (the detection with the highest score) and no FAs. This causes that any other systems working at different miss/FA ratios have a better DET curve in case there are FAs with better scores than those of the hits.

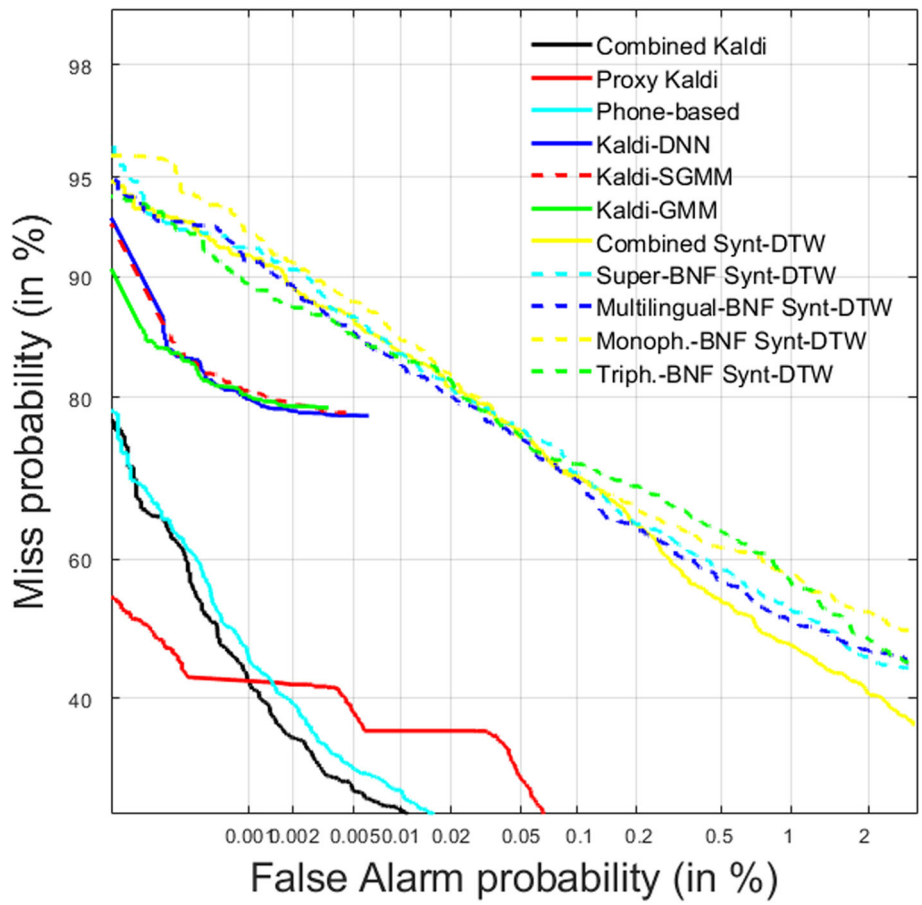

Fig. 16 The DET curves of the STD systems for RTVE test data

5 Post-evaluation analysis

After the evaluation period, an analysis based on some term properties and fusion of the primary systems submitted from the different participants has been carried out. This section presents the results of this analysis.

5.1 Performance analysis of STD systems for in-language and out-of-language terms

An analysis of the STD performance has been carried out for INL and OOL terms and results are presented in Fig. 18 for MAVIR, RTVE, and COREMAH test data. As expected, a large STD performance degradation is found from INL to OOL terms for all the databases in systems that employ a text-based STD approach. Some discrepancies appear on the COREMAH test data (i.e., *Kaldi-DNN*, *Kaldi-SGMM*, and *Kaldi-GMM* systems) although the STD performance is so low that any conclusion is meaningless. However, for the systems that employ the QbE-STD approach, the differences between INL and OOL terms are not so clear, and specially on MAVIR data, systems perform, in general, better for OOL term detection than for INL term detection. This may be due to the fact that Spanish language was not employed in the feature extraction, but English and other IARPA Babel program languages were, along with the fact that OOL terms in this database are mainly English terms.

5.2 Performance analysis of STD systems for single and multi-word terms

An analysis of the STD performance has been carried out for single and multi-word terms and results are presented in Fig. 19 for MAVIR, RTVE, and COREMAH test data. They show some differences depending on the database.

On MAVIR test data, *Kaldi-DNN*, *Kaldi-SGMM*, *Kaldi-GMM*, and *Proxy Kaldi* systems perform better for single-word term detection than for multi-word term detection. This probably happens because multi-word term detection is intrinsically more difficult for word-based ASR since more words must be detected. However, on RTVE test data, *Kaldi-DNN*, *Kaldi-SGMM*, and *Kaldi-GMM* systems perform better for multi-word term detection than for single-word term detection. This might be caused by the fact that there are much more OOV single-word

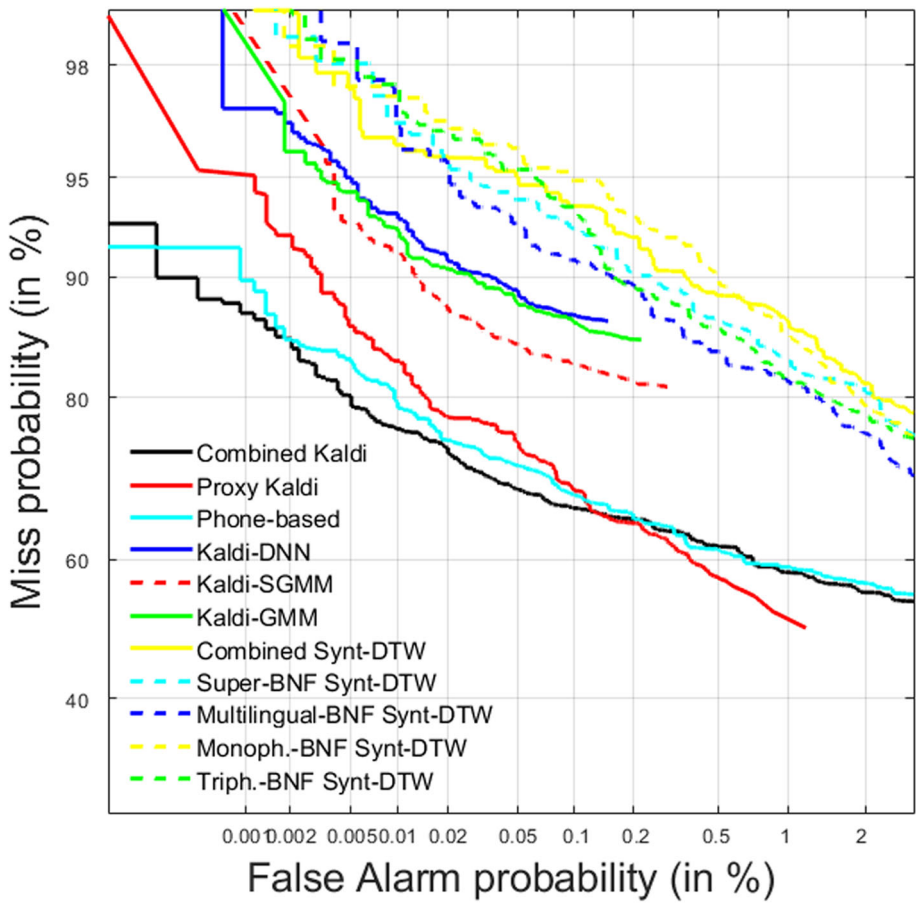

Fig. 17 The DET curves of the STD systems for COREMAH test data

terms (202) than OOV multi-word terms (9), which leads to a dramatical degradation in the final performance for single-word terms.

On the other hand, *Combined Kaldi* and *Phone-based* systems perform better for multi-word term detection on MAVIR and RTVE test data. The *Phone-based* system relies on a subword unit approach and multi-word terms are typically longer than single-word terms. Short terms tend to produce many FAs in phone-based systems, and the opposite stands for longer terms. Therefore, phone-based systems may obtain better performance for multi-word term detection. The *Combined Kaldi* system performance for multi-word terms seems to be highly influenced by the *Phone-based* system.

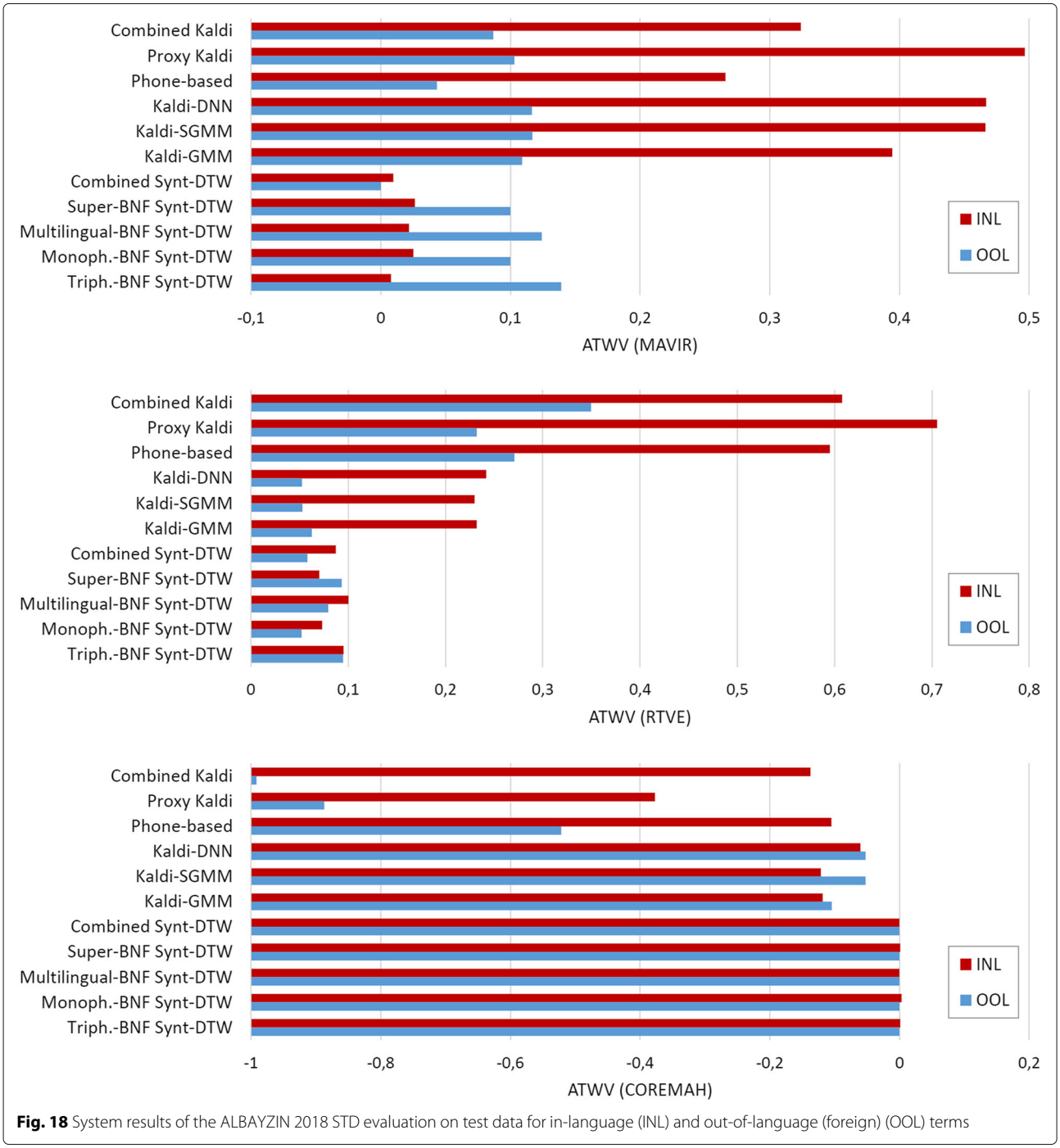

Fig. 18 System results of the ALBAYZIN 2018 STD evaluation on test data for in-language (INL) and out-of-language (foreign) (OOL) terms

The systems that employ a QbE-STD approach obtain better performance for multi-word term detection than for single-word term detection on MAVIR and RTVE test data due to the fact that multi-word terms are typically longer than single-word terms so that less false alarms are produced with the DTW search. In addition, these systems typically perform better for multi-word term detection than the *Kaldi-DNN*, *Kaldi-SGMM*, and *Kaldi-GMM* systems. This indicates that QbE-STD approaches can be effectively employed for long term detection in the absence of robust word-based LVCSR systems.

On COREMAH test data, the systems obtained better performance for multi-word term detection than for single-word term detection. However, the ATWVs are, in general, so low that any conclusion is hardly reliable.

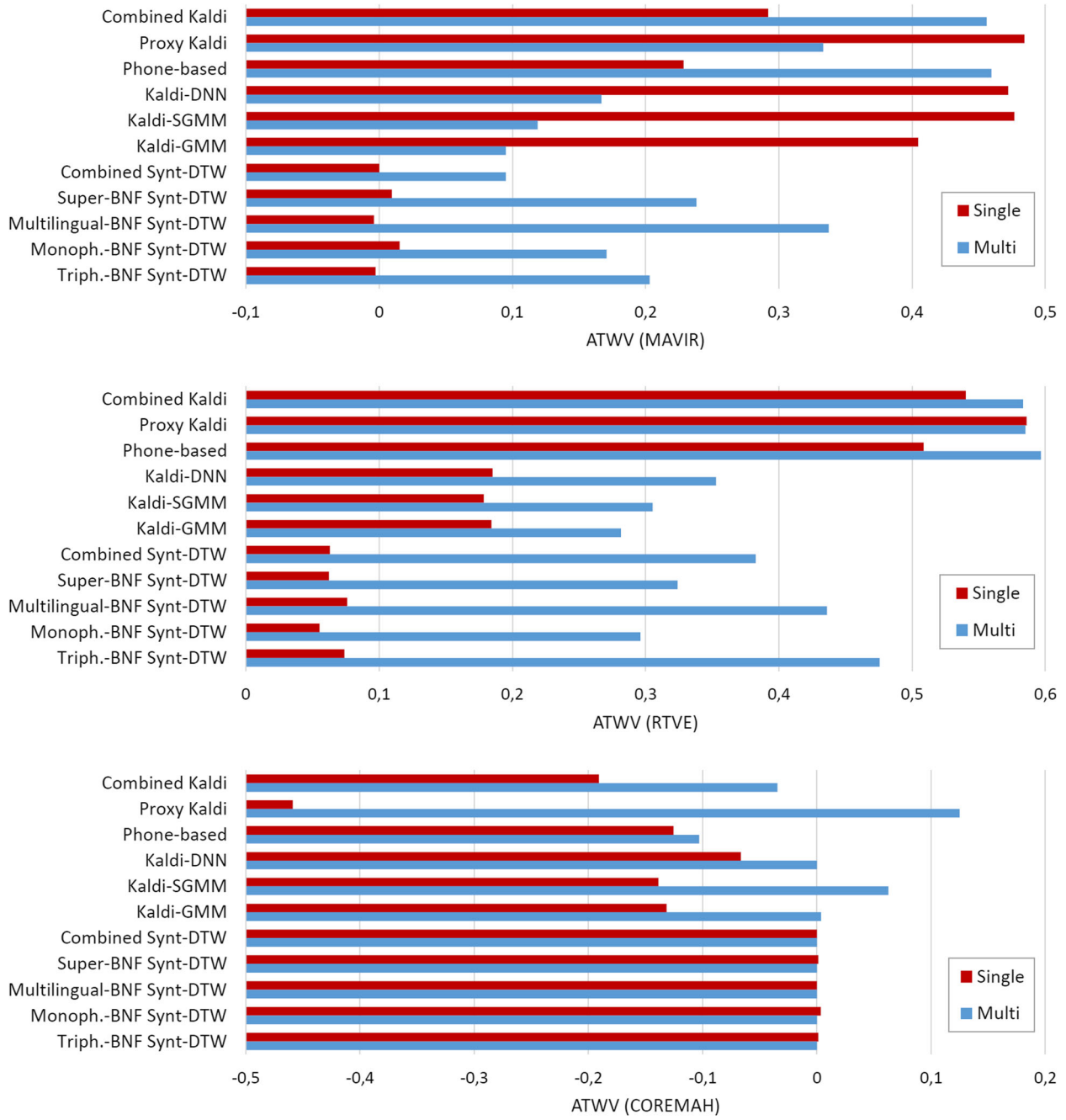

Fig. 19 System results of the ALBAYZIN 2018 STD evaluation on test data for single-word (Single) and multi-word (Multi) terms

5.3 Performance analysis of STD systems for INV and OOV terms

Figure 20 shows a similar analysis for in-vocabulary and out-of-vocabulary terms. The text-based STD systems perform, in general, better for INV term detection than for OOV term detection. The *Kaldi-DNN*, *Kaldi-SGMM*, and *Kaldi-GMM* systems on MAVIR test data are the only exceptions. We consider this could be due to the moderate OOV term rate (20.3%) in this dataset, along with the amount of training data used to train the INV language model. However, when the OOV term rate increases (66.8% for the RTVE test data), the proxy words strategy of Kaldi for OOV term detection is less powerful. System performance is so low on

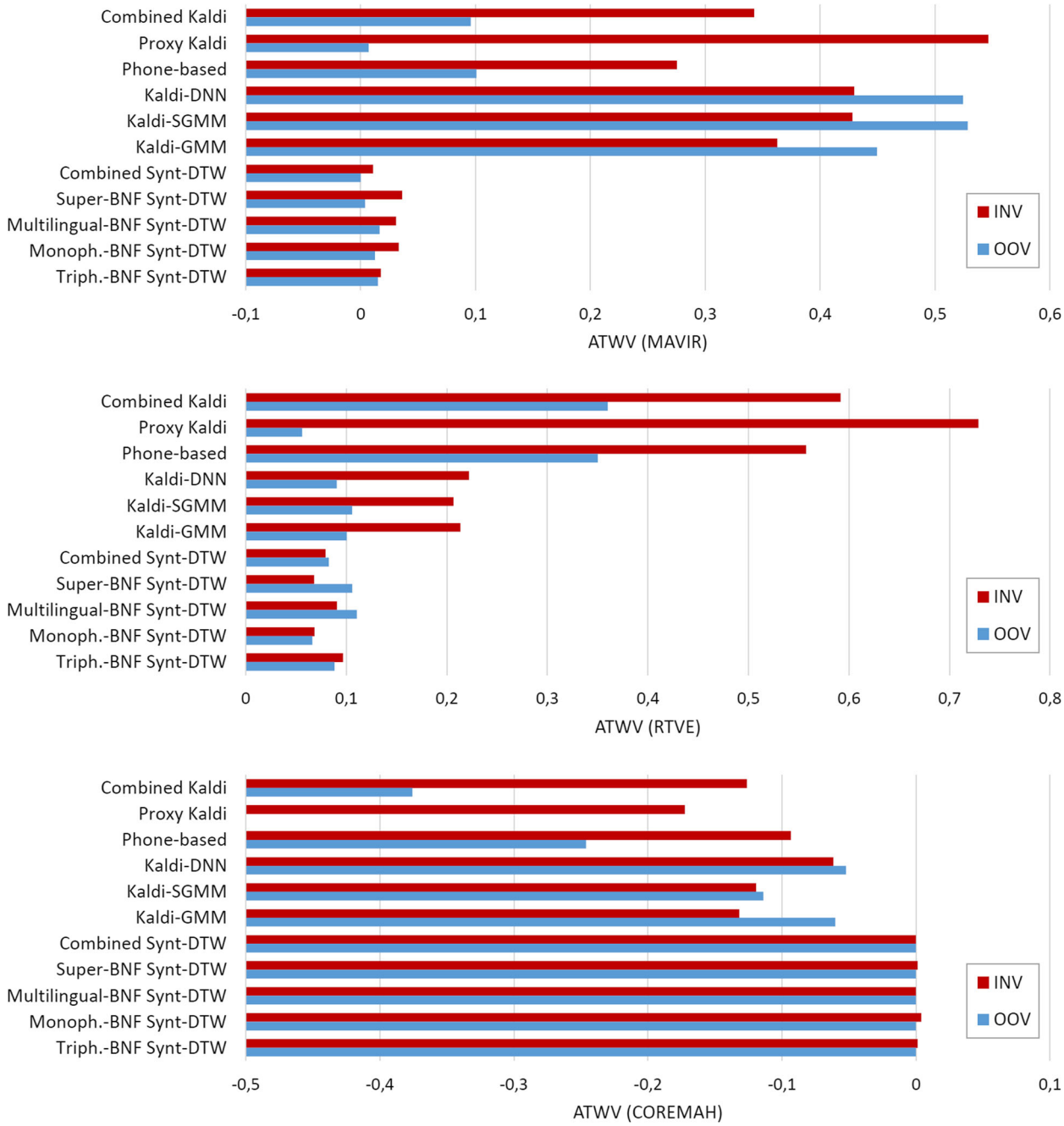

Fig. 20 System results of the ALBAYZIN 2018 STD evaluation on test data for in-vocabulary (INV) and out-of-vocabulary (OOV) terms

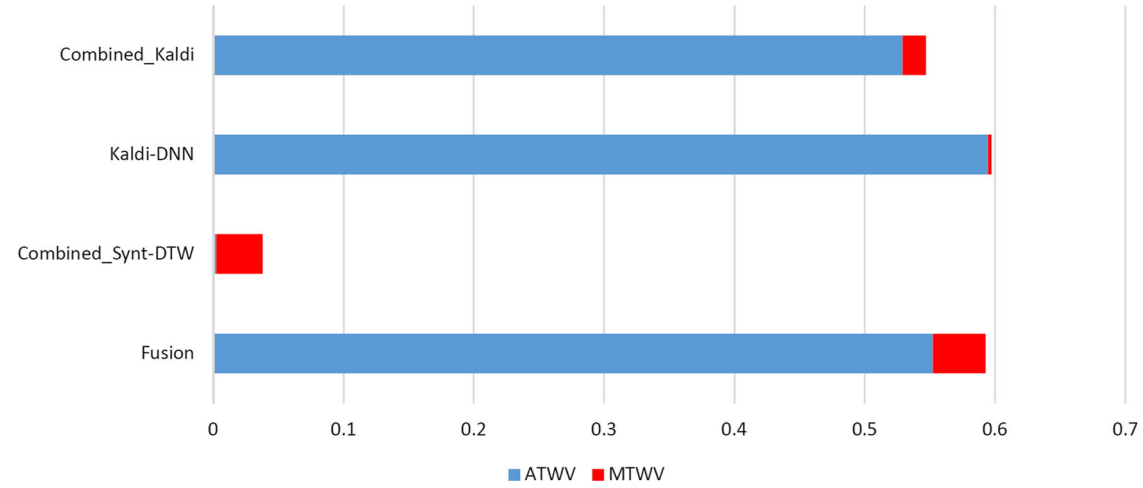

Fig. 21 Fusion and primary system results of the ALBAYZIN 2018 STD evaluation on MAVIR development data

COREMAH test data that no reliable conclusion can be derived.

5.4 System fusion

After the evaluation, we have combined all the primary systems developed by the participants by fusing the scores they produced. System fusion consists of two different stages: (1) pre-processing and (2) calibration and fusion. These are explained next.

5.4.1 Pre-processing

First, scores for each query and system are normalized to mean 0 and variance 1.

All the detections given by the fused systems are taken into account to generate the output of the *fusion* system. Given a certain detection output by a certain system A, in case some other fused system B does not detect it (and hence there is no score for it), the score generated for that detection is the minimum global score for all the terms generated by system B.

5.4.2 Calibration and fusion

Calibration and fusion are carried out with the Bosaris toolkit [130]. To do so, a linear model based on logistic regression trained on the development detection scores is employed. MAVIR and RTVE fusion parameters are optimized independently based on their corresponding development sets and then, are applied to their corresponding test sets. For COREMAH data, the model trained for MAVIR data is employed.

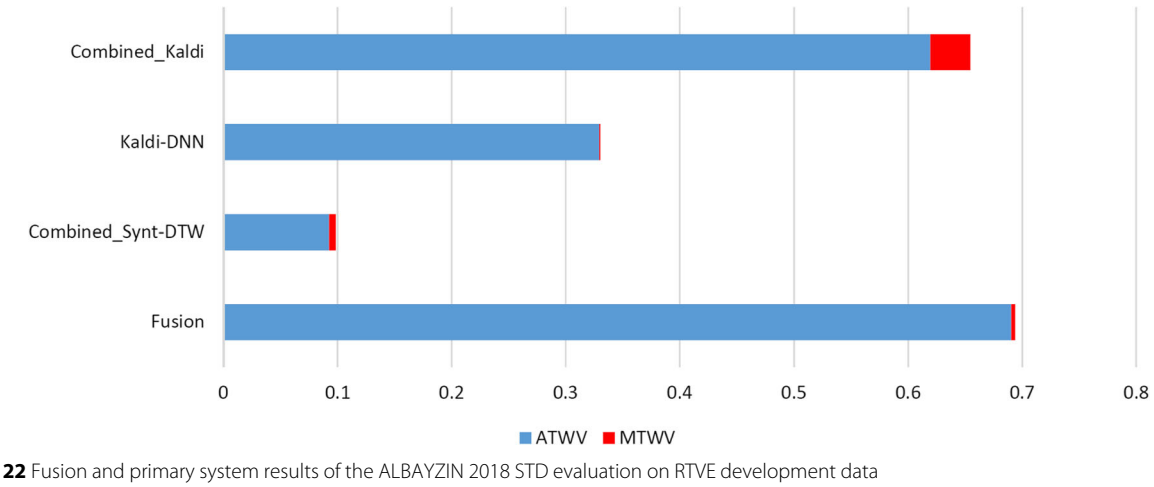

Fig. 22 Fusion and primary system results of the ALBAYZIN 2018 STD evaluation on RTVE development data

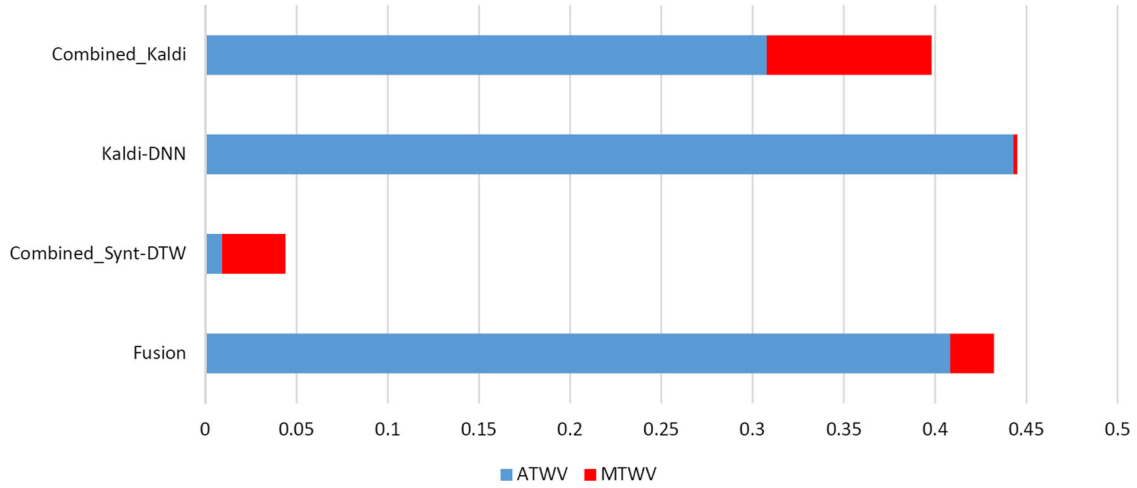

Fig. 23 Fusion and primary system results of the ALBAYZIN 2018 STD evaluation on MAVIR test data

Fusion is employed to combine the three primary systems of the participants in the evaluation (i.e., *Combined Kaldi*, *Combined Synt-DTW*, and *Kaldi-DNN* systems).

5.4.3 Fusion results

Results are presented in Figs. 21 and 22 for MAVIR and RTVE development data, respectively, and in Figs. 23, 24, and 25 for MAVIR, RTVE, and COREMAH test data, respectively. They show that system fusion plays an important role on RTVE data, for which the fusion improves the best individual system for both development and test data. A paired t test shows that the *Fusion* system obtains a statistically significant difference ($p < 0.01$) for both sets of RTVE data. However, on MAVIR and COREMAH data, the fusion does not outperform the best individual system. RTVE data contain higher-quality/better-pronounced speech than MAVIR data, and there were much more data available for RTVE. Fusion gets more benefit on these conditions. On COREMAH data, for which there are no available data for a fine tuning, fusion gets also worse results.

DET curves of the primary systems and the fusion systems are presented in Figs. 26 and 27 for MAVIR and RTVE development data, respectively, and in Figs. 28, 29, and 30 for MAVIR, RTVE, and COREMAH test data, respectively. On MAVIR development data, the *Combined Kaldi* system performs the best for low FA rates, the *Kaldi-DNN* system performs the best for moderate FA

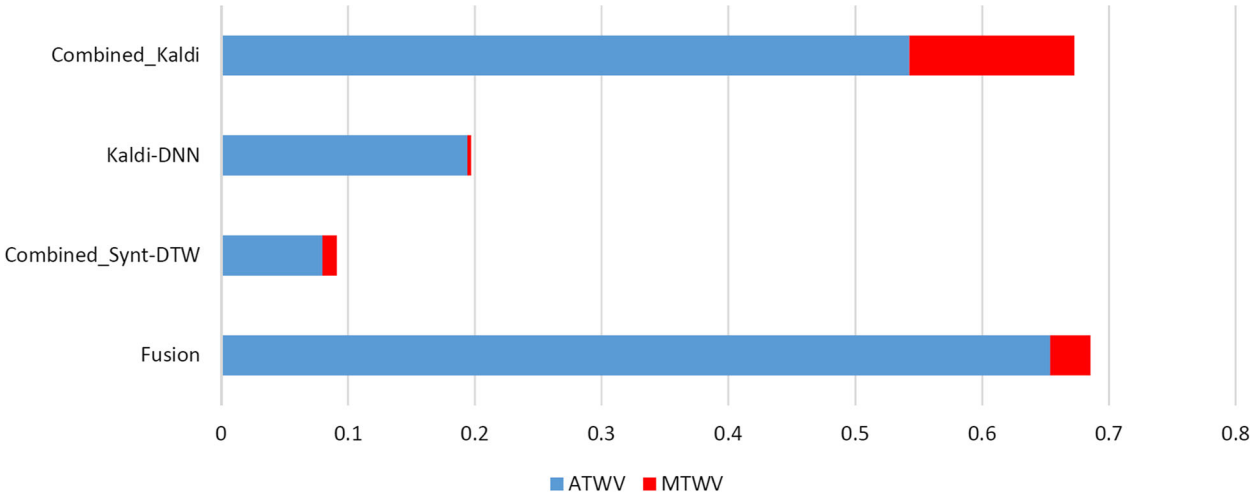

Fig. 24 Fusion and primary system results of the ALBAYZIN 2018 STD evaluation on RTVE test data

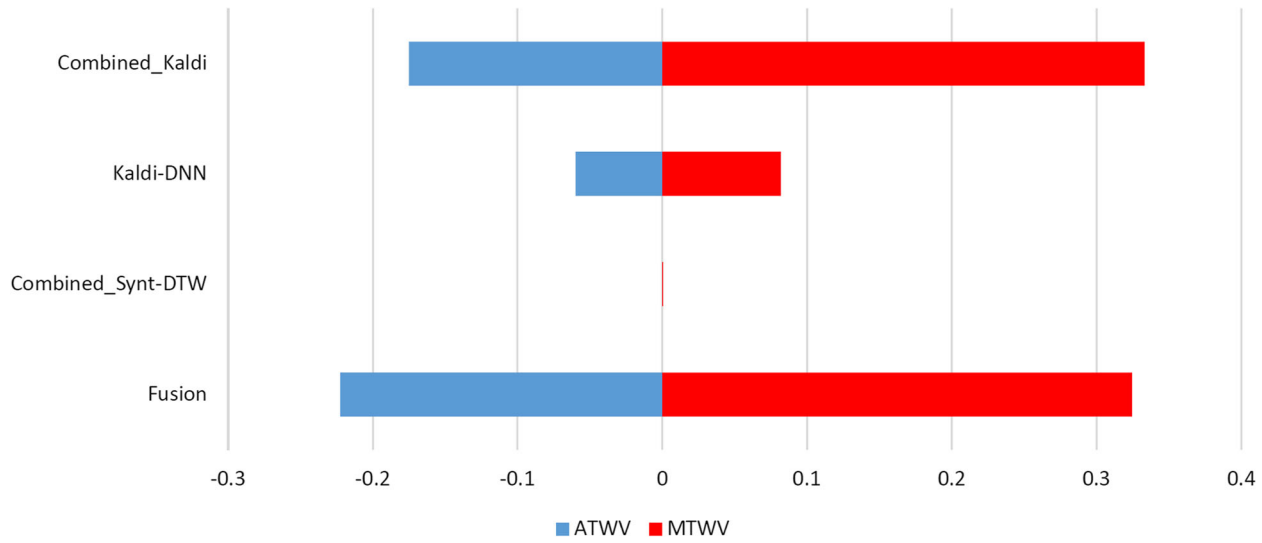

Fig. 25 Fusion and primary system results of the ALBAYZIN 2018 STD evaluation on COREMAH test data

rates, and the *Fusion* system performs the best for low miss rates. This means that the fusion may be suitable for scenarios in which misses are more important than false alarms. On MAVIR test data, the *Kaldi-DNN* system performs the best for low FA rates and the *Fusion* system performs the best for low miss rates. This confirms that system fusion is suitable for low miss rates on MAVIR data. On RTVE development and test data, the *Fusion* system performs the best for almost all the operating points, which is consistent with the MTWV/ATWV results (see

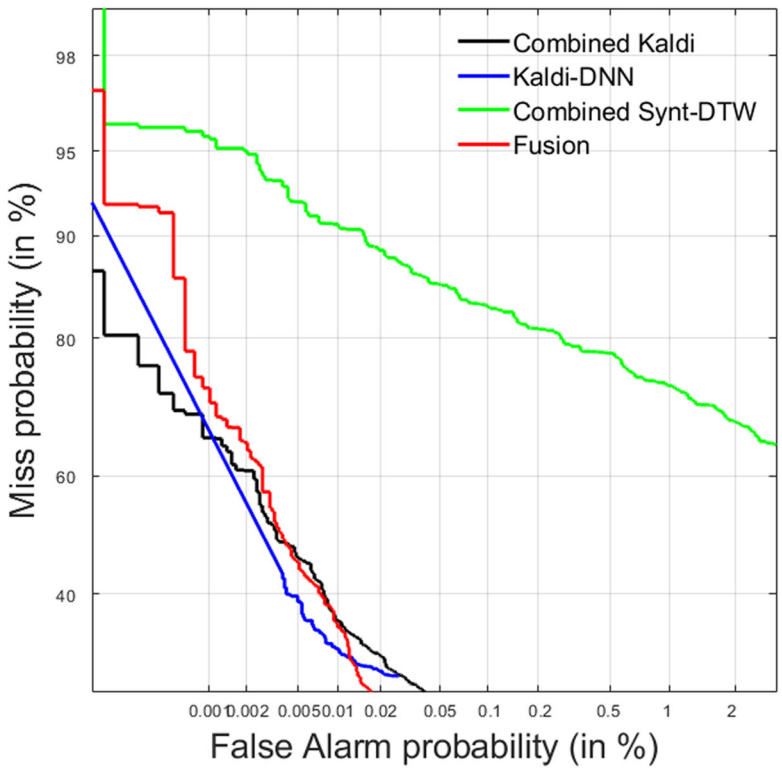

Fig. 26 The DET curves of the fusion and primary STD systems for MAVIR development data

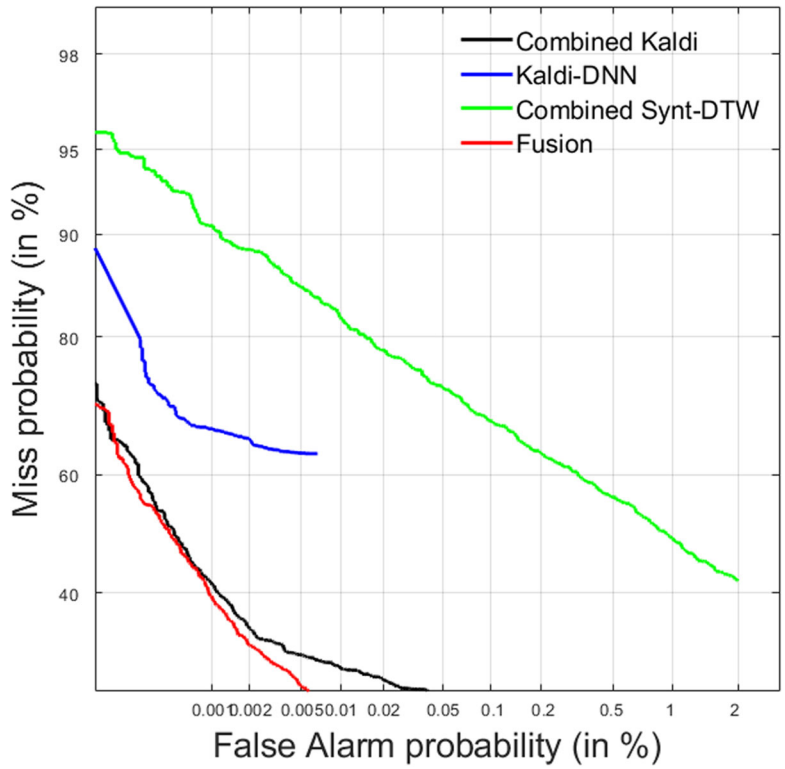

Fig. 27 The DET curves of the fusion and primary STD systems for RTVE development data

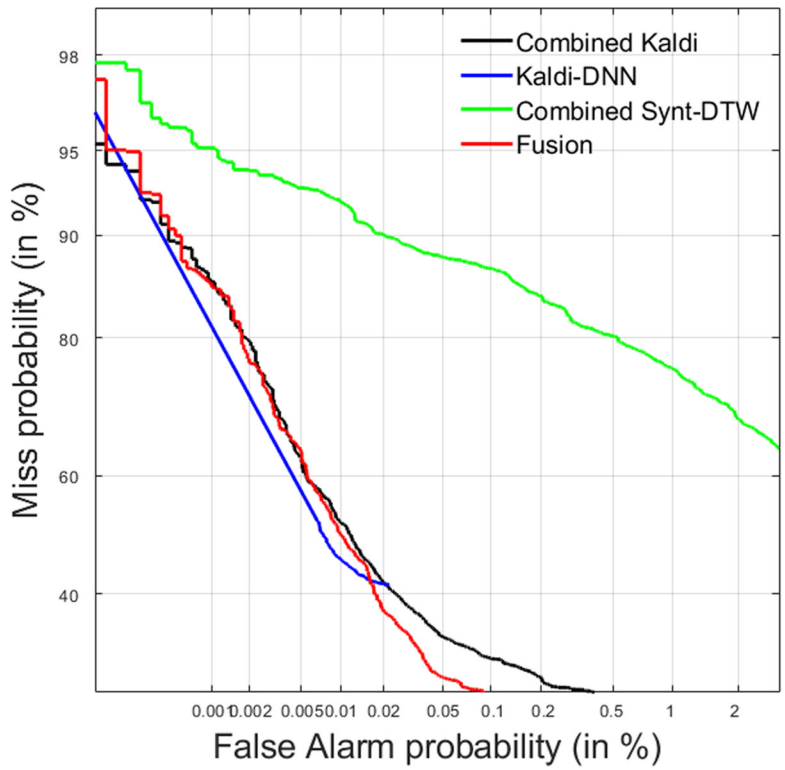

Fig. 28 The DET curves of the fusion and primary STD systems for MAVIR test data

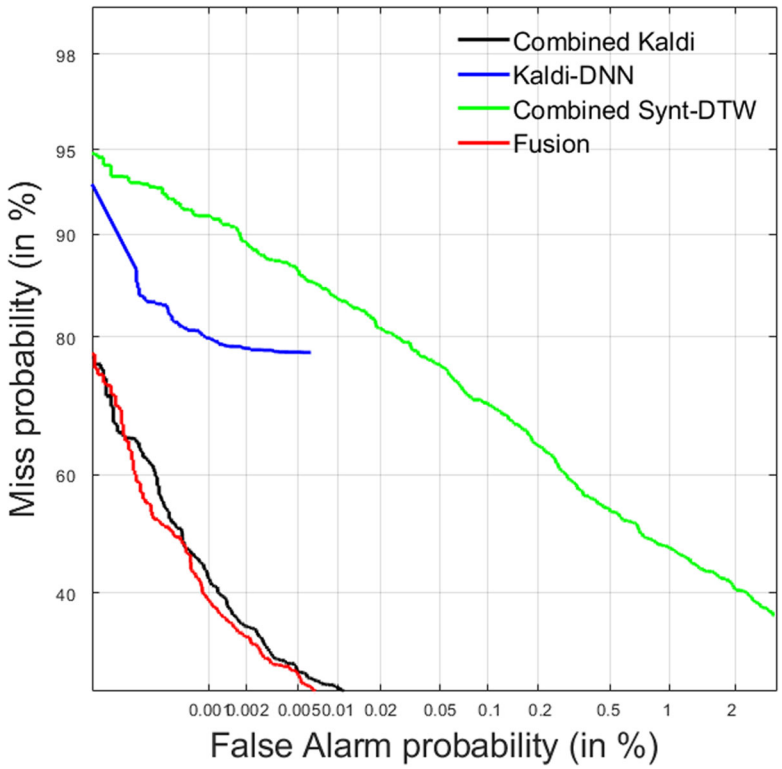

Fig. 29 The DET curves of the fusion and primary STD systems for RTVE test data

Figs. 22 and 24). On COREMAH test data, the *Combined Kaldi* system performs the best for almost all the operating points. In these data, the *Fusion* system may play an important role for low miss rates regardless the high FA rates.

5.5 Comparison to the ALBAYZIN 2016 STD evaluation

Given that MAVIR test data are the same for ALBAYZIN 2016 and 2018 STD evaluations, these data can be used to establish a comparison between the performance of the systems presented to both evaluations (see Fig. 31). The highest performance in 2018 (ATWV = 0.4699) is lower than that obtained in 2016 (ATWV = 0.5724). In the 2018 evaluation, the decision threshold was tuned on MAVIR and RTVE data simultaneously, which produced the performance degradation. However, in 2016, the decision threshold was only tuned on MAVIR data, which produced a better threshold calibration, and hence, better performance.

Therefore, it can be said that building multi-domain STD systems still represents a research challenge since it can lead to reduced performance on some specific domains. However, this presents a great advantage, since a single system is able to search on speech in different domains.

6 Conclusions

This paper has presented a multi-domain spoken term detection international 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. Three different research groups have taken part in the evaluation and eleven different systems were submitted in total. Most of the systems are largely based on the standard text-based STD approach (with state-of-the-art DNN-based ASR systems) for hypothesizing detections from word ASR. On the other hand, other systems are based on a QbE-STD framework for hypothesizing detections. Among those systems, *Combined Kaldi* and *Phone-based* systems, which include a probabilistic retrieval model for information retrieval and a query likelihood retrieval model, and *Combined Synt-DTW*, *Super-BNF Synt-DTW*, *Multilingual-BNF Synt-DTW*, *Monoph.-BNF Synt-DTW*, and *Triph.-BNF Synt-DTW*, which employ speech synthesis for query generation from the term list and a QbE-STD approach, can be considered novel from an STD perspective.

The most important conclusion from this evaluation is that multi-domain STD is still a challenge in STD research, since results have shown much variability with

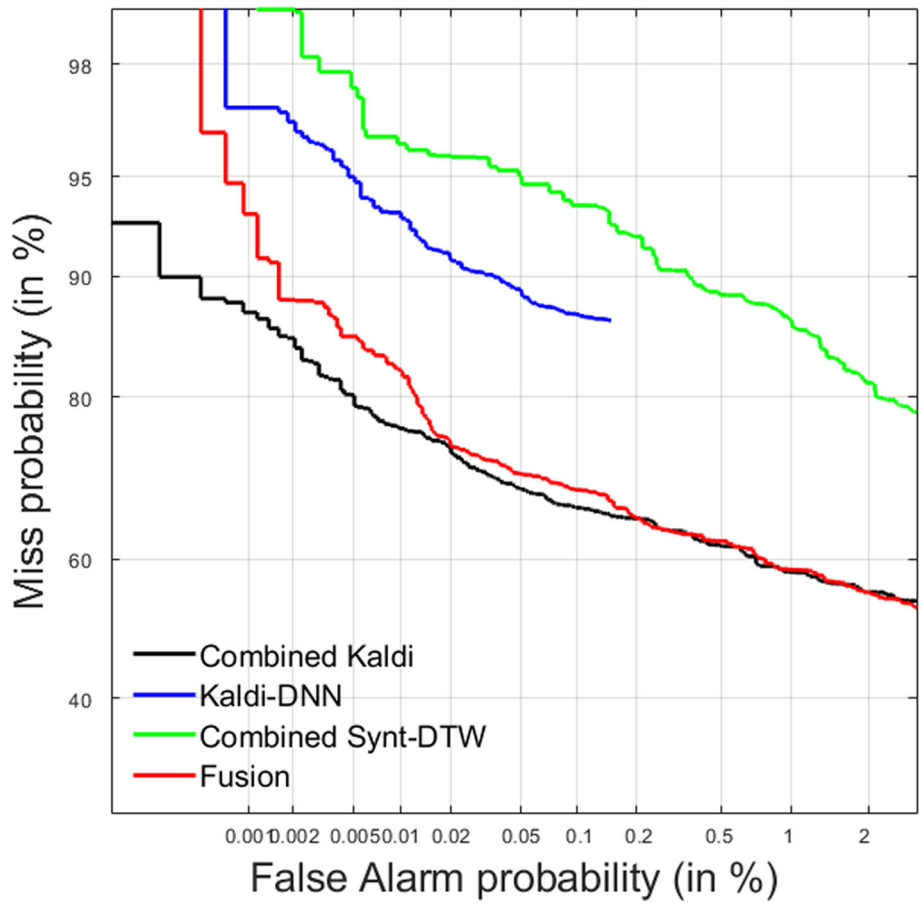

Fig. 30 The DET curves of the fusion and primary STD systems for COREMAH test data

regard to domain mismatch. On the one hand, the submitted systems have obtained the best performance on RTVE data, for which more data are available for system construction and include high-quality and well-pronounced speech. On the other hand, the systems have obtained the worst performance on COREMAH data, for which only test data were provided, and speech is very spontaneous and with high degree of overlapping. This indicates that domain change is quite challenging in STD tasks. Finally, systems on MAVIR data, which present highly-spontaneous speech, obtained performances between those obtained on RTVE and COREMAH data.

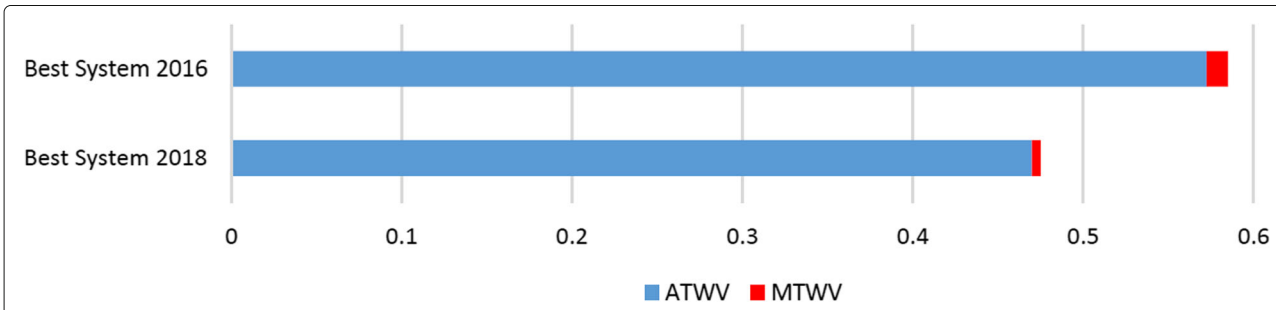

Fig. 31 Best system performances obtained in the ALBAYZIN STD 2016 and 2018 evaluations on MAVIR test data

We have also shown that OOL term detection still remains an important challenge in STD, since systems have obtained low performance on those terms (ATWV = 0.1392 on MAVIR data and ATWV = 0.3496 on RTVE data). On OOV term detection, which is crucial for open-vocabulary STD, systems have obtained best performances of ATWV = 0.5284 on MAVIR data and ATWV = 0.3600 on RTVE data (i.e., for domains in which training/development data have been provided). Regarding multi-word term detection, systems have obtained best performances of ATWV = 0.4595 on MAVIR data and ATWV = 0.5967 on RTVE data.

Given the best overall result obtained in the evaluation (ATWV = 0.2250), which comes from the average of the three domains, there is still ample room for improvement. Specifically, the performance of STD systems degrades dramatically when applied to *unseen data*. This encourages us to maintain the STD evaluation in the next years, focusing on multi-domain STD, and the applicability of this technology to unseen challenging domains. Specifically, in the next months we will be launching the ALBAYZIN 2020 STD evaluation to be held in November 2020 within the IberSPEECH conference. This new evaluation edition aims to provide new domains and more challenging data (i.e., more difficult search terms) and evaluation conditions (i.e., rank the submitted systems from weighting the system performance according to the most challenging domain).

Appendix

This appendix shows the full result tables for the systems submitted to the ALBAYZIN 2018 STD evaluation for development and test data.

Table 12 Overall system results of the ALBAYZIN 2018 STD evaluation on development and test data

	Development		Test	
System ID	MTWV	ATWV	MTWV	ATWV
Combined Kaldi	0.6001	0.5743	0.4098	0.2250
Proxy Kaldi	0.5645	0.5489	0.3825	0.2187
Phone-based	0.5557	0.5366	0.3723	0.2135
Kaldi-DNN	0.4639	0.4621	0.2213	0.1924
Kaldi-SGMM	0.4467	0.4383	0.2200	0.1698
Kaldi-GMM	0.4324	0.4248	0.2062	0.1489
Combined Synt-DTW	0.0683	0.0475	0.0452	0.0297
Super-BNF Synt-DTW	0.0691	0.0565	0.0427	0.0362
Multilingual-BNF Synt-DTW	0.0737	0.0640	0.0433	0.0412
Monoph.-BNF Synt-DTW	0.0665	0.0565	0.0369	0.0338
Triph.-BNF Synt-DTW	0.0655	0.0492	0.0451	0.0378

Table 13 System results of the ALBAYZIN 2018 STD evaluation on MAVIR development data

System ID	MTWV	ATWV	*p*(FA)	*p*(Miss)
Combined Kaldi	0.5470	0.5290	0.00011	0.348
Proxy Kaldi	0.5314	0.5179	0.00003	0.442
Phone-based	0.4828	0.4739	0.00016	0.362
Kaldi-DNN	0.5974	0.5946	0.00008	0.322
Kaldi-SGMM	0.6045	0.5897	0.00005	0.348
Kaldi-GMM	0.5705	0.5556	0.00006	0.373
Combined Synt-DTW	0.0379	0.0023	0.00000	0.961
Super-BNF Synt-DTW	0.0469	0.0293	0.00004	0.917
Multilingual-BNF Synt-DTW	0.0467	0.0351	0.00001	0.947
Monoph.-BNF Synt-DTW	0.0332	0.0191	0.00002	0.950
Triph.-BNF Synt-DTW	0.0405	0.0137	0.00001	0.952

Table 14 System results of the ALBAYZIN 2018 STD evaluation on RTVE development data

System ID	MTWV	ATWV	*p*(FA)	*p*(Miss)
Combined Kaldi	0.6549	0.6195	0.00004	0.306
Proxy Kaldi	0.5976	0.5798	0.00001	0.397
Phone-based	0.6286	0.5993	0.00004	0.331
Kaldi-DNN	0.3303	0.3295	0.00002	0.648
Kaldi-SGMM	0.2889	0.2868	0.00001	0.699
Kaldi-GMM	0.2943	0.2939	0.00002	0.690
Combined Synt-DTW	0.0986	0.0927	0.00004	0.859
Super-BNF Synt-DTW	0.0912	0.0836	0.00002	0.888
Multilingual-BNF Synt-DTW	0.1007	0.0928	0.00002	0.878
Monoph.-BNF Synt-DTW	0.0997	0.0939	0.00002	0.880
Triph.-BNF Synt-DTW	0.0905	0.0846	0.00005	0.862

Table 15 System results of the ALBAYZIN 2018 STD evaluation on MAVIR test data

System ID	MTWV	ATWV	*p*(FA)	*p*(Miss)
Combined Kaldi	0.3981	0.3077	0.00014	0.464
Proxy Kaldi	0.4750	0.4699	0.00011	0.412
Phone-based	0.3432	0.2506	0.00015	0.506
Kaldi-DNN	0.4450	0.4429	0.00009	0.464
Kaldi-SGMM	0.4478	0.4424	0.00008	0.475
Kaldi-GMM	0.4046	0.3750	0.00008	0.510
Combined Synt-DTW	0.0440	0.0091	0.00002	0.941
Super-BNF Synt-DTW	0.0365	0.0315	0.00001	0.953
Multilingual-BNF Synt-DTW	0.0343	0.0288	0.00002	0.943
Monoph.-BNF Synt-DTW	0.0309	0.0303	0.00001	0.963
Triph.-BNF Synt-DTW	0.0317	0.0171	0.00001	0.963

Table 16 System results of the ALBAYZIN 2018 STD evaluation on RTVE test data

System ID	MTWV	ATWV	*p*(FA)	*p*(Miss)
Combined Kaldi	0.6730	0.5425	0.00005	0.277
Proxy Kaldi	0.5860	0.5859	0.00006	0.355
Phone-based	0.6458	0.5133	0.00005	0.304
Kaldi-DNN	0.1970	0.1941	0.00001	0.789
Kaldi-SGMM	0.1876	0.1851	0.00002	0.797
Kaldi-GMM	0.1916	0.1894	0.00002	0.793
Combined Synt-DTW	0.0911	0.0799	0.00003	0.880
Super-BNF Synt-DTW	0.0864	0.0761	0.00004	0.878
Multilingual-BNF Synt-DTW	0.0957	0.0949	0.00004	0.869
Monoph.-BNF Synt-DTW	0.0746	0.0679	0.00003	0.897
Triph.-BNF Synt-DTW	0.1022	0.0951	0.00002	0.881

Table 17 System results of the ALBAYZIN 2018 STD evaluation on COREMAH test data

System ID	MTWV	ATWV	*p*(FA)	*p*(Miss)
Combined Kaldi	0.1582	− 0.1751	0.00005	0.791
Proxy Kaldi	0.0864	− 0.3997	0.00005	0.860
Phone-based	0.1279	− 0.1233	0.00002	0.849
Kaldi-DNN	0.0220	− 0.0598	0.00001	0.971
Kaldi-SGMM	0.0247	− 0.1182	0.00004	0.931
Kaldi-GMM	0.0224	− 0.1178	0.00002	0.959
Combined Synt-DTW	0.0006	0.0000	0.00001	0.988
Super-BNF Synt-DTW	0.0051	0.0011	0.00000	0.993
Multilingual-BNF Synt-DTW	0.0000	0.0000	0.00000	1.000
Monoph.-BNF Synt-DTW	0.0053	0.0032	0.00000	0.995
Triph.-BNF Synt-DTW	0.0014	0.0011	0.00000	0.999

Table 18 Fusion and primary system results of the ALBAYZIN 2018 STD evaluation on development data

	MAVIR				RTVE			
System ID	MTWV	ATWV	*p*(FA)	*p*(Miss)	MTWV	ATWV	*p*(FA)	*p*(Miss)
Combined Kaldi	0.5470	0.5290	0.00011	0.348	0.6549	0.6195	0.00004	0.306
Kaldi-DNN	0.5974	0.5946	0.00008	0.322	0.3303	0.3295	0.00002	0.648
Combined Synt-DTW	0.0379	0.0023	0.00000	0.961	0.0986	0.0927	0.00004	0.859
Fusion	0.5928	0.5523	0.00015	0.261	0.6940	0.6903	0.00006	0.242

Table 19 Fusion and primary system results of the ALBAYZIN 2018 STD evaluation on test data

	MAVIR				RTVE			
System ID	MTWV	ATWV	*p*(FA)	*p*(Miss)	MTWV	ATWV	*p*(FA)	*p*(Miss)
Combined Kaldi	0.3981	0.3077	0.00014	0.464	0.6730	0.5425	0.00005	0.277
Kaldi-DNN	0.4450	0.4429	0.00009	0.464	0.1970	0.1941	0.00001	0.789
Combined Synt-DTW	0.0440	0.0091	0.00002	0.941	0.0911	0.0799	0.00003	0.880
Fusion	0.4323	0.4084	0.00019	0.376	0.6854	0.6539	0.00005	0.261
	COREMAH							
Combined Kaldi	0.1582	− 0.1751	0.00005	0.791				
Kaldi-DNN	0.0220	− 0.0598	0.00001	0.971				
Combined Synt-DTW	0.0006	0.0000	0.00001	0.988				
Fusion	0.1021	− 0.2225	0.00004	0.856				

Abbreviations
AAC: Advanced audio coding; ASR: Automatic speech recognition; ATWV: Actual term-weighted value; BNews: Broadcast news; CLI: Command-line interface; CMVN: Cepstral mean and variance normalization; CTS: Conversational telephone speech; DET: Detection error tradeoff; DNN: Deep neural network; DTW: Dynamic time warping; FA: False alarm; fMLLR: Feature-space maximum likelihood linear regression; GMMs: Gaussian mixture models; gTTS: Google TTS; HMM: Hidden markov model; HTK: Hidden Markov model toolkit; INL: In-language; INV: In-vocabulary; KWS: Keyword spotting; LDA: Linear discriminant analysis; LM: Language model; LVCSR: Large vocabulary continuous speech recognition; MED: Minimum edit distance; MFCCs: Mel-frequency cepstral coefficients; MLLT: Maximum likelihood linear transform; MOS: Mean opinion score; MPEG: Moving picture experts group; MTWV: Maximum term-weighted value; NIST: National institute of Standards and Technology; OOL: Out-of-language; OOV: Out-of-vocabulary; OpenKWS: Open keyword search; OpenSAT: Open speech analytics technologies; PCM: Pulse code modulation; QbE: Query-by-example; RTMeet: Roundtable meeting rooms; RTTH: Spanish thematic network on speech technologies; RTVE: Radio Televisión Española; sBNF: Stacked bottleneck feature; SDR: Spoken document retrieval; SGMM: Subspace Gaussian mixture model; SIG-IL: ISCA Special interest group on iberian languages; SoS: Search on speech; STD: Spoken term detection; TST: Term-specific threshold; TTS: Text-to-speech; TV: Television; TWV: Term-weighted value; VAD: Voice activity detection; WFSTs: Weighted finite state transducers; WSJ: Wall Street Journal; XML: Extensible Markup Language

Acknowledgements
Not applicable.

Authors' contributions
JT and DTT designed, prepared the STD evaluation, and carried out the post-evaluation analysis. PL-O and LD-F built the *Combined Kaldi*, *Proxy Kaldi*, and *Phone-based* systems. Ana RM and JRR built the *Kaldi-DNN*, *Kaldi-SGMM*, and *Kaldi-GMM* systems. MP and LJ R-F built the *Combined Synt-DTW*, *Super-BNF Synt-DTW*, *Multilingual-BNF Synt-DTW*, *Monoph.-BNF Synt-DTW*, and *Triph.-BNF Synt-DTW* systems and carried out the primary system fusion. All the authors contributed in the final discussion of the results. The main contributions of this paper are the following: (1) Systems submitted to the Fourth Spoken Term Detection Evaluation for Spanish Language are presented. (2) New challenging database based on Spanish broadcast news has been used. (3) Analysis of system results and primary system fusion for the three different domains are presented. All authors read and approved the final manuscript.

Authors' information
Not applicable.

Funding
This work has received financial support from "Ministerio de Economía y Competitividad" of the Government of Spain, the European Regional Development Fund (ERDF) under the research projects TIN2015-64282-R, RTI2018-093336-B-C22 TEC2015-65345-P, Xunta de Galicia (projects GPC ED431B 2016/035, GPC ED431B 2019/003, and GRC 2014/024), Xunta de Galicia - "Consellería de Cultura, Educación e Ordenación Universitaria", the ERDF through the 2016–2019 accreditations ED431G/01 ("Centro singular de investigación de Galicia") and ED431G/04 ("Agrupación estratéxica consolidada"), the UPV/EHU under grant GIU16/68, and the project "DSSL: Redes Profundas y Modelos de Subespacios para Detección y Seguimiento de Locutor, Idioma y Enfermedades Degenerativas a partir de la Voz" (TEC2015-68172-C2-1-P, MINECO/FEDER).

Availability of data and materials
Please contact author for data requests.

Competing interests
The authors declare that they have no competing interests.

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]Voice Group, Advanced Technologies Application Center (CENATAV), Rpto. Siboney, Playa, Havana, Cuba. [6]Software Technology Working Group (GTTS), Universidad del País Vasco, Barrio Sarriena s/n, Leioa, Spain.

Received: 7 March 2019 Accepted: 22 July 2019
Published online: 02 September 2019

References
1. M. Larson, G. Jones, Spoken content retrieval: a survey of techniques and technologies. Found. Trends Inf. Retr. **5**(4-5), 235–422 (2011)
2. K. Ng, V. W. Zue, Subword-based approaches for spoken document retrieval. Speech Comm. **32**(3), 157–186 (2000)
3. B. Chen, K.-Y. Chen, P.-N. Chen, Y.-W. Chen, Spoken document retrieval with unsupervised query modeling techniques. IEEE Trans. Audio, Speech, Lang. Process. **20**(9), 2602–12 (2012)
4. T.-H. Lo, Y.-W. Chen, K.-Y. Chen, H.-M. Wang, B. Chen, in *Proceedings of ASRU*. Neural relevance-aware query modeling for spoken document retrieval (IEEE, USA, 2017), pp. 466–473
5. W. F. L. Heeren, F. M. G. de Jong, L. B. van der Werff, M. A. H. Huijbregts, R. J. F. Ordelman, in *Proceedings of LREC*. Evaluation of spoken document retrieval for historic speech collections (ELRA, Belgium, 2008), pp. 2037–2041
6. Y.-C. Pan, H.-Y. Lee, L.-S. Lee, Interactive spoken document retrieval with suggested key terms ranked by a Markov decision process. IEEE Trans Audio, Speech, Lang. Process. **20**(2), 632–645 (2012)
7. Y.-W. Chen, K.-Y. Chen, H.-M. Wang, B. Chen, in *Proceedings of Interspeech*. Exploring the use of significant words language modeling for spoken document retrieval (ISCA, France, 2017), pp. 2889–2893
8. P. Gao, J. Liang, P. Ding, B. Xu, in *Proceedings of ICASSP*. A novel phone-state matrix based vocabulary-independent keyword spotting method for spontaneous speech (IEEE, USA, 2007), pp. 425–428
9. B. Zhang, R. Schwartz, S. Tsakalidis, L. Nguyen, S. Matsoukas, in *Proceedings of Interspeech*. White listing and score normalization for keyword spotting of noisy speech (ISCA, France, 2012), pp. 1832–1835
10. A. Mandal, J. van Hout, Y.-C. Tam, V. Mitra, Y. Lei, J. Zheng, D. Vergyri, L. Ferrer, M. Graciarena, A. Kathol, H. Franco, in *Proceedings of Interspeech*. Strategies for high accuracy keyword detection in noisy channels (ISCA, France, 2013), pp. 15–19
11. T. Ng, R. Hsiao, L. Zhang, D. Karakos, S. H. Mallidi, M. Karafiat, K. Vesely, I. Szoke, B. Zhang, L. Nguyen, R. Schwartz, in *Proceedings of Interspeech*. Progress in the BBN keyword search system for the DARPA RATS program (ISCA, France, 2014), pp. 959–963
12. V. Mitra, J. van Hout, H. Franco, D. Vergyri, Y. Lei, M. Graciarena, Y.-C. Tam, J. Zheng, in *Proceedings of ICASSP*. Feature fusion for high-accuracy keyword spotting (IEEE, USA, 2014), pp. 7143–7147
13. S. Panchapagesan, M. Sun, A. Khare, S. Matsoukas, A. Mandal, B. Hoffmeister, S. Vitaladevuni, in *Proceedings of Interspeech*. Multi-task learning and weighted cross-entropy for DNN-based keyword spotting (ISCA, France, 2016), pp. 760–764
14. J. Mamou, B. Ramabhadran, O. Siohan, in *Proceedings of ACM SIGIR*. Vocabulary independent spoken term detection (ACM, USA, 2007), pp. 615–622
15. D. Schneider, T. Mertens, M. Larson, J. Kohler, in *Proceedings of Interspeech*. Contextual verification for open vocabulary spoken term detection (ISCA, France, 2010), pp. 697–700
16. C. Parada, A. Sethy, M. Dredze, F. Jelinek, in *Proceedings of Interspeech*. A spoken term detection framework for recovering out-of-vocabulary words using the web (ISCA, France, 2010), pp. 1269–1272
17. I. Szöke, M. Fapšo, L. Burget, J. Černocký, in *Proceedings of Speech Search Workshop at SIGIR*. Hybrid word-subword decoding for spoken term detection (ACM, USA, 2008), pp. 42–48
18. Y. Wang, F. Metze, in *Proceedings of Interspeech*. An in-depth comparison of keyword specific thresholding and sum-to-one score normalization (ISCA, France, 2014), pp. 2474–2478
19. L. Mangu, G. Saon, M. Picheny, B. Kingsbury, in *Proceedings of ICASSP*. Order-free spoken term detection (IEEE, USA, 2015), pp. 5331–5335
20. A. Buzo, H. Cucu, C. Burileanu, in *Proceedings of MediaEval*. SpeeD@MediaEval 2014: spoken term detection with robust multilingual phone recognition (CEUR, Germany, 2014), pp. 721–722
21. R. Konno, K. Ouchi, M. Obara, Y. Shimizu, T. Chiba, T. Hirota, Y. Itoh, in *Proceedings of NTCIR-12*. An STD system using multiple STD results and

multiple rescoring method for NTCIR-12 SpokenQuery&Doc task (Japan Society for Promotion of Science, Japan, 2016), pp. 200–204
22. R. Jarina, M. Kuba, R. Gubka, M. Chmulik, M. Paralic, in *Proceedings of MediaEval*. UNIZA system for the spoken web search task at MediaEval 2013 (CEUR, Germany, 2013), pp. 791–792
23. X. Anguera, M. Ferrarons, in *Proceedings of ICME*. Memory efficient subsequence DTW for query-by-example spoken term detection (IEEE, USA, 2013)
24. H. Lin, A. Stupakov, J. Bilmes, in *Proceedings of Interspeech*. Spoken keyword spotting via multi-lattice alignment (ISCA, France, 2008), pp. 2191–2194
25. C. Chan, L. Lee, in *Proceedings of Interspeech*. Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping (ISCA, France, 2010), pp. 693–696
26. J. Mamou, B. Ramabhadran, in *Proceedings of Interspeech*. Phonetic query expansion for spoken document retrieval (ISCA, France, 2008), pp. 2106–2109
27. D. Can, E. Cooper, A. Sethy, C. White, B. Ramabhadran, M. Saraclar, in *Proceedings of ICASSP*. Effect of pronunciations on OOV queries in spoken term detection (IEEE, USA, 2009), pp. 3957–3960
28. A. Rosenberg, K. Audhkhasi, A. Sethy, B. Ramabhadran, M. Picheny, in *Proceedings of ICASSP*. End-to-end speech recognition and keyword search on low-resource languages (IEEE, USA, 2017), pp. 5280–5284
29. K. Audhkhasi, A. Rosenberg, A. Sethy, B. Ramabhadran, B. Kingsbury, in *Proceedings of ICASSP*. End-to-end ASR-free keyword search from speech (IEEE, USA, 2017), pp. 4840–4844
30. K. Audhkhasi, A. Rosenberg, A. Sethy, B. Ramabhadran, B. Kingsbury, End-to-end ASR-free keyword search from speech. IEEE J. Sel. Topics Signal Process. **11**(8), 1351–1359 (2017)
31. J. G. Fiscus, J. Ajot, J. S. Garofolo, G. Doddingtion, in *Proceedings of Workshop on Searching Spontaneous Conversational Speech*. Results of the 2006 spoken term detection evaluation (ACM, USA, 2007), pp. 45–50
32. W. Hartmann, L. Zhang, K. Barnes, R. Hsiao, S. Tsakalidis, R. Schwartz, in *Proceedings of Interspeech*. Comparison of multiple system combination techniques for keyword spotting (ISCA, France, 2016), pp. 1913–1917
33. T. Alumae, D. Karakos, W. Hartmann, R. Hsiao, L. Zhang, L. Nguyen, S. Tsakalidis, R. Schwartz, in *Proceedings of ICASSP*. The 2016 BBN Georgian telephone speech keyword spotting system (IEEE, USA, 2017), pp. 5755–5759
34. D. Vergyri, A. Stolcke, R. R. Gadde, W. Wang, in *Proceedings of NIST Spoken Term Detection Workshop (STD 2006)*. The SRI 2006 spoken term detection system (NIST, USA, 2006), pp. 1–15
35. D. Vergyri, I. Shafran, A. Stolcke, R. R. Gadde, M. Akbacak, B. Roark, W. Wang, in *Proceedings of Interspeech*. The SRI/OGI 2006 spoken term detection system (ISCA, France, 2007), pp. 2393–2396
36. M. Akbacak, D. Vergyri, A. Stolcke, in *Proceedings of ICASSP*. Open-vocabulary spoken term detection using graphone-based hybrid recognition systems (IEEE, USA, 2008), pp. 5240–5243
37. I. Szöke, M. Fapšo, M. Karafiát, L. Burget, F. Grézl, P. Schwarz, O. Glembek, P. Matějka, J. Kopecký, J. Černocký, in *Machine Learning for Multimodal Interaction*. Spoken term detection system based on combination of LVCSR and phonetic search, vol. 4892 (Springer, Germany, 2008), pp. 237–247
38. I. Szöke, L. Burget, J. Černocký, M. Fapšo, in *Proceedings of SLT*. Sub-word modeling of out of vocabulary words in spoken term detection (IEEE, USA, 2008), pp. 273–276
39. S. Meng, P. Yu, J. Liu, F. Seide, in *Proceedings of ICASSP*. Fusing multiple systems into a compact lattice index for Chinese spoken term detection (IEEE, USA, 2008), pp. 4345–4348
40. K. Thambiratmann, S. Sridharan, Rapid yet accurate speech indexing using dynamic match lattice spotting. IEEE Trans. Audio, Speech, Lang. Process. **15**(1), 346–357 (2007)
41. R. Wallace, R. Vogt, B. Baker, S. Sridharan, in *Proceedings of ICASSP*. Optimising figure of merit for phonetic spoken term detection (IEEE, USA, 2010), pp. 5298–5301
42. A. Jansen, K. Church, H. Hermansky, in *Proceedings of Interspeech*. Towards spoken term discovery at scale with zero resources (ISCA, France, 2010), pp. 1676–1679
43. C. Parada, A. Sethy, B. Ramabhadran, in *Proceedings of ICASSP*. Balancing false alarms and hits in spoken term detection (IEEE, USA, 2010), pp. 5286–5289
44. J. Trmal, M. Wiesner, V. Peddinti, X. Zhang, P. Ghahremani, Y. Wang, V. Manohar, H. Xu, D. Povey, S. Khudanpur, in *Proceedings of Interspeech*. The Kaldi OpenKWS system: i low resource keyword search (ISCA, France, 2017), pp. 3597–3601
45. C.-A. Chan, L.-S. Lee, in *Proceedings of Interspeech*. Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping (ISCA, France, 2010), pp. 693–696
46. C.-P. Chen, H.-Y. Lee, C.-F. Yeh, L.-S. Lee, in *Proceedings of Interspeech*. Improved spoken term detection by feature space pseudo-relevance feedback (ISCA, France, 2010), pp. 1672–1675
47. P. Motlicek, F. Valente, P. Garner, in *Proceedings of Interspeech*. English spoken term detection in multilingual recordings (ISCA, France, 2010), pp. 206–209
48. I. Szöke, M. Fapšo, M. Karafiát, L. Burget, F. Grézl, P. Schwarz, O. Glembek, P. Matějka, S. Kontár, J. Černocký, in *Proceedings of NIST Spoken Term Detection Evaluation Workshop (STD'06)*. BUT system for NIST STD 2006 - English (NIST, USA, 2006), pp. 1–15
49. D. R. H. Miller, M. Kleber, C.-L. Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, H. Gish, in *Proceedings of Interspeech*. Rapid and accurate spoken term detection (ISCA, France, 2007), pp. 314–317
50. H. Li, J. Han, T. Zheng, G. Zheng, in *Proceedings of Interspeech*. A novel confidence measure based on context consistency for spoken term detection (ISCA, France, 2012), pp. 2430–2433
51. J. Chiu, A. Rudnicky, in *Proceedings of Interspeech*. Using conversational word bursts in spoken term detection (ISCA, France, 2013), pp. 2247–2251
52. C. Ni, C.-C. Leung, L. Wang, N. F. Chen, B. Ma, in *Proceedings of ICASSP*. Efficient methods to train multilingual bottleneck feature extractors for low resource keyword search (IEEE, USA, 2017), pp. 5650–5654
53. Z. Meng, B.-H. Juang, in *Proceedings of Interspeech*. Non-uniform boosted MCE training of deep neural networks for keyword spotting (ISCA, France, 2016), pp. 770–774
54. Z. Meng, B.-H. Juang, in *Proceedings of Interspeech*. Non-uniform MCE training of deep long short-term memory recurrent neural networks for keyword spotting (ISCA, France, 2017), pp. 3547–3551
55. S.-w. Lee, K. Tanaka, Y. Itoh, in *Proceedings of Interspeech*. Combination of diverse subword units in spoken term detection (ISCA, France, 2015), pp. 3685–3289
56. C. van Heerden, D. Karakos, K. Narasimhan, M. Davel, R. Schwartz, in *Proceedings of ICASSP*. Constructing sub-word units for spoken term detection (IEEE, USA, 2017), pp. 5780–5784
57. D. Kaneko, R. Konno, K. Kojima, K. Tanaka, S.-w. Lee, Y. Itoh, in *Proceedings of Interspeech*. Constructing acoustic distances between subwords and states obtained from a deep neural network for spoken term detection (ISCA, France, 2017), pp. 2879–2883
58. V. T. Pham, H. Xu, X. Xiao, N. F. Chen, E. S. Chng, in *Proceedings of International Symposium on Information and Communication Technology*. Pruning strategies for partial search in spoken term detection (ACM, USA, 2017), pp. 114–119
59. M. Wollmer, B. Schuller, G. Rigoll, Keyword spotting exploiting long short-term memory. Speech Comm. **55**(2), 252–265 (2013)
60. J. Tejedor, D. T. Toledano, D. Wang, S. King, J. Colás, Feature analysis for discriminative confidence estimation in spoken term detection. Comput. Speech Lang. **28**(5), 1083–1114 (2014). Elsevier, Amsterdam
61. Y. Zhuang, X. Chang, Y. Qian, K. Yu, in *Proceedings of Interspeech*. Unrestricted vocabulary keyword spotting using LSTM-CTC (ISCA, France, 2016), pp. 938–942
62. L. Pandey, K. Nathwani, in *Proceedings of Interspeech*. LSTM based attentive fusion of spectral and prosodic information for keyword spotting in hindi language (ISCA, France, 2018), pp. 112–116
63. R. Lileikyte, T. Fraga-Silva, L. Lamel, J.-L. Gauvain, A. Laurent, G. Huang, in *Proceedings of ICASSP*. Effective keyword search for low-resourced conversational speech (IEEE, USA, 2017), pp. 5785–5789
64. S. Parlak, M. Saraçlar, in *Proceedings of ICASSP*. Spoken term detection for Turkish broadcast news (IEEE, USA, 2008), pp. 5244–5247
65. V. T. Pham, H. Xu, X. Xiao, N. F. Chen, E. S. Chng, Re-ranking spoken term detection with acoustic exemplars of keywords. Speech Comm. **104**, 12–23 (2018)
66. A. Ragni, D. Saunders, P. Zahemszky, J. Vasilakes, M. J. F. Gales, K. M. Knill, in *Proceedings of ICASSP*. Morph-to-word transduction for accurate and efficient automatic speech recognition and keyword search (IEEE, USA, 2017), pp. 5770–5774
67. X. Chen, A. Ragnil, J. Vasilakes, X. Liu, K. Knilll, M. J. . F. Gales, in *Proceedings of ICASSP*. Recurrent neural network language models for keyword search (IEEE, USA, 2017), pp. 5775–5779

68. D. Xu, F. Metze, in *Proceedings of Interspeech*. Word-based probabilistic phonetic retrieval for low-resource spoken term detection (ISCA, France, 2014), pp. 2774–2778
69. J. Svec, J. V. Psutka, L. Smidl, J. Trmal, in *Proceedings of Interspeech*. A relevance score estimation for spoken term detection based on RNN-generated pronunciation embeddings (ISCA, France, 2017), pp. 2934–2938
70. Y. Khokhlov, I. Medennikov, A. Romanenko, V. Mendelev, M. Korenevsky, A. Prudnikov, N. Tomashenko, A. Zatvornitsky, in *Proceedings of Interspeech*. The STC keyword search system for OpenKWS 2016 evaluation (ISCA, France, 2017), pp. 3602–3606
71. S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK Book (v3.4). Engineering Department, Cambridge University (2009)
72. K.-F. Lee, H.-W. Hon, R. Reddy, An overview of the SPHINX speech recognition system. IEEE Trans. Acoust., Speech, Signal Process. **38**(1), 35–45 (1990)
73. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, in *Proceedings of ASRU*. The KALDI speech recognition toolkit (IEEE, USA, 2011)
74. G. Chen, S. Khudanpur, D. Povey, J. Trmal, D. Yarowsky, O. Yilmaz, in *Proceedings of ICASSP*. Quantifying the value of pronunciation lexicons for keyword search in low resource languages (IEEE, USA, 2013), pp. 8560–8564
75. V. T. Pham, N. F. Chen, S. Sivadas, H. Xu, I.-F. Chen, C. Ni, E. S. Chng, H. Li, in *Proceedings of SLT*. System and keyword dependent fusion for spoken term detection (IEEE, USA, 2014), pp. 430–435
76. G. Chen, O. Yilmaz, J. Trmal, D. Povey, S. Khudanpur, in *Proceedings of ASRU*. Using proxies for OOV keywords in the keyword search task (IEEE, USA, 2013), pp. 416–421
77. B. Taras, C. Nadeu, Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion. EURASIP J. Audio, Speech, Music Process. **1**, 1–10 (2011)
78. M. Zelenák, H. Schulz, J. Hernando, Speaker diarization of broadcast news in Albayzin 2010 evaluation campaign. EURASIP J. Audio, Speech, Music Process. **19**, 1–9 (2012)
79. L. J. Rodríguez-Fuentes, M. Penagarikano, A. Varona, M. Díez, G. Bordel, in *Proceedings of Interspeech*. The Albayzin 2010 Language Recognition Evaluation (ISCA, France, 2011), pp. 1529–1532
80. J. Tejedor, D. T. Toledano, P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, A. Cardenal, J. D. Echeverry-Correa, A. Coucheiro-Limeres, J. Olcoz, A. Miguel, Spoken term detection ALBAYZIN 2014 evaluation: overview, systems, results, and discussion. EURASIP, J. Audio, Speech Music Process. **2015**(21), 1–27 (2015)
81. J. Tejedor, D. T. Toledano, X. Anguera, A. Varona, L. F. Hurtado, A. Miguel, J. Colás, Query-by-example spoken term detection ALBAYZIN 2012 evaluation: overview, systems, results, and discussion. EURASIP, J. Audio, Speech Music Process. **2013**(23), 1–17 (2013)
82. J. Tejedor, D. T. Toledano, P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, Comparison of ALBAYZIN query-by-example spoken term detection 2012 and 2014 evaluations. EURASIP, J. Audio, Speech Music Process. **2016**(1), 1–19 (2016)
83. D. Castán, D. Tavarez, P. Lopez-Otero, J. Franco-Pedroso, H. Delgado, E. Navas, L. Docio-Fernández, D. Ramos, J. Serrano, A. Ortega, E. Lleida, Albayzín-2014 evaluation: audio segmentation and classification in broadcast news domains. EURASIP, J. Audio, Speech Music Process. **2015**(33), 1–9 (2015)
84. F. Méndez, L. Docío, M. Arza, F. Campillo, in *Proceedings of FALA*. The Albayzin 2010 text-to-speech evaluation (ISCA, France, 2010), pp. 317–340
85. J. Tejedor, D. T. Toledano, P. Lopez-Otero, L. Docio-Fernandez, L. Serrano, I. Hernaez, A. Coucheiro-Limeres, J. Ferreiros, J. Olcoz, J. Llombart, Albayzin 2016 spoken term detection evaluation: an international open competitive evaluation in spanish. EURASIP, J. Audio, Speech Music Process. **2017**(22), 1–23 (2017)
86. J. Tejedor, D. T. Toledano, P. Lopez-Otero, L. Docio-Fernandez, J. Proença, F. P. ao, F. García-Granada, E. Sanchis, A. Pompili, A. Abad, Albayzin query-by-example spoken term detection 2016 evaluation. EURASIP, J. Audio, Speech Music Process. **2018**(2), 1–25 (2018)
87. J. Billa, K. W. Ma, J. W. McDonough, Zavaliagkos, D. R. Miller, K. N. Ross, A. El-Jaroudi, in *Proceedings of Eurospeech*. Multilingual speech recognition: the 1996 Byblos callhome system (ISCA, France, 1997)
88. H. Cuayahuitl, B. Serridge, in *Proceedings of MICAI*. Out-of-vocabulary word modeling and rejection for spanish keyword spotting systems (Springer, Germany, 2002), pp. 156–165
89. M. Killer, S. Stuker, T. Schultz, in *Proceedings of Eurospeech*. Grapheme based speech recognition (ISCA, France, 2003), pp. 3141–3144
90. J. Tejedor, *Contributions to Keyword Spotting and Spoken Term Detection For Information Retrieval in Audio Mining PhD thesis*. (Universidad Autónoma de Madrid, Madrid, Spain, 2009)
91. L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, D. Povey, A. Rastrow, R. C. Rose, S. Thomas, in *Proceedings of ICASSP*. Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models (IEEE, USA, 2010), pp. 4334–4337
92. J. Tejedor, D. T. Toledano, D. Wang, S. King, J. Colás, Feature analysis for discriminative confidence estimation in spoken term detection. Comput. Speech Lang. **28**(5), 1083–1114 (2014)
93. J. Li, X. Wang, B. Xu, in *Proceedings of Interspeech*. An empirical study of multilingual and low-resource spoken term detection using deep neural networks (ISCA, France, 2014), pp. 1747–1751
94. M. Hazewinkel, *Student Test*. (Kluwer Academic, Denmark, 1994)
95. NIST, *The Spoken Term Detection (STD) 2006 Evaluation Plan*, 10th edn. (National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 2006). http://www.nist.gov/speech/tests/std
96. A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, in *Proceedings of Eurospeech*. The DET curve in assessment of detection task performance (ISCA, France, 1997), pp. 1895–1898
97. NIST, *Evaluation Toolkit (STDEval) Software*. (National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 1996). http://www.itl.nist.gov/iad/mig/tests/std/tools
98. 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 11 Aug 2019
99. E. Lleida, A. Ortega, A. Miguel, V. Bazán, C. Pérez, M. Zotano, A. de Prada, *RTVE2018 Database Description*. (Vivolab and Corporación Radiotelevisión Española, Zaragoza, Spain, 2018). http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf
100. M. V. Matos, Diseño y compilación de un corpus multimodal de análisis pragmático para la aplicación a la enseñanza del español. PhD thesis (Universidad Autónoma de Madrid, Madrid, 2017)
101. Z. Lv, M. Cai, W.-Q. Zhang, J. Liu, in *Proceedings of Interspeech*. A novel discriminative score calibration method for keyword search (ISCA, France, 2016), pp. 745–749
102. W. Hartmann, L. Zhang, K. Barnes, R. Hsiao, S. Tsakalidis, R. Schwartz, in *Proceedings of Interspeech*. Comparison of multiple system combination techniques for keyword spotting (ISCA, France, 2016), pp. 1913–1917
103. N. F. Chen, V. T. Pharri, H. Xu, X. Xiao, V. H. Do, C. Ni, I.-F. Chen, S. Sivadas, C.-H. Lee, E. S. Chng, B. Ma, H. Li, in *Proceedings of ICASSP*. Exemplar-inspired strategies for low-resource spoken keyword search in Swahili (IEEE, USA, 2016), pp. 6040–6044
104. C. Ni, C.-C. Leung, L. Wang, H. Liu, F. Rao, L. Lu, N. F. Chen, B. Ma, H. Li, in *Proceedings of ICASSP*. Cross-lingual deep neural network based submodular unbiased data selection for low-resource keyword search (IEEE, USA, 2016), pp. 6015–6019
105. M. Cai, Z. Lv, C. Lu, J. Kang, L. Hui, Z. Zhang, J. Liu, in *Proceedings of ASRU*. High-performance swahili keyword search with very limited language pack: The THUEE system for the OpenKWS15 evaluation (IEEE, USA, 2015), pp. 215–222
106. N. F. Chen, C. Ni, I.-F. Chen, S. Sivadas, V. T. Pham, H. Xu, X. Xiao, T. S. Lau, S. J. Leow, B. P. Lim, C.-C. Leung, L. Wang, C.-H. Lee, A. Goh, E. S. Chng, B. Ma, H. Li, in *Proceedings of ICASSP*. Low-resource keyword search strategies for Tamil (IEEE, USA, 2015), pp. 5366–5370
107. W. Hartmann, D. Karakos, R. Hsiao, L. Zhang, T. Alumae, S. Tsakalidis, R. Schwartz, in *Proceedings of ICASSP*. Analysis of keyword spotting performance across IARPA babel languages (IEEE, USA, 2017), pp. 5765–5769
108. NIST, *OpenKWS13 Keyword Search Evaluation Plan*. (National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 2013). https://www.nist.gov/sites/default/files/documents/itl/iad/mig/OpenKWS13-evalplan-v4.pdf
109. NIST, *Draft KWS14 Keyword Search Evaluation Plan*. (National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 2013). https://

www.nist.gov/sites/default/files/documents/itl/iad/mig/KWS14-evalplan-v11.pdf
110. NIST, *KWS15 Keyword Search Evaluation Plan*. (National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 2015). https://www.nist.gov/sites/default/files/documents/itl/iad/mig/KWS15-evalplan-v05.pdf
111. NIST, *Draft KWS16 Keyword Search Evaluation Plan*. (National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 2016). https://www.nist.gov/sites/default/files/documents/itl/iad/mig/KWS16-evalplan-v04.pdf
112. T. Akiba, H. Nishizaki, K. Aikawa, T. Kawahara, T. Matsui, in *Proceedings of NTCIR-9*. Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop (Japan Society for Promotion of Science, Japan, 2011), pp. 1–13
113. T. Akiba, H. Nishizaki, K. Aikawa, X. Hu, Y. Itoh, T. Kawahara, S. Nakagawa, H. Nanjo, Y. Yamashita, in *Proceedings of NTCIR-10*. Overview of the NTCIR-10 SpokenDoc-2 task (Japan Society for Promotion of Science, Japan, 2013), pp. 1–15
114. T. Akiba, H. Nishizaki, H. Nanjo, G. J. F. Jones, in *Proceedings of NTCIR-11*. Overview of the NTCIR-11 SpokenQuery&Doc Task (Japan Society for Promotion of Science, Japan, 2014), pp. 1–15
115. T. Akiba, H. Nishizaki, H. Nanjo, G. J. F. Jones, in *Proceedings of NTCIR-12*. Overview of the NTCIR-12 SpokenQuery&Doc-2 Task (Japan Society for Promotion of Science, Japan, 2016), pp. 1–13
116. K. Vesely, A. Ghoshal, L. Burget, D. Povey, in *Proceedings of Interspeech*. Sequence-discriminative training of deep neural networks (ISCA, France, 2013), pp. 2345–2349
117. P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, S. Khudanpur, in *Proceedings of ICASSP*. A pitch extraction algorithm tuned for automatic speech recognition (IEEE, USA, 2014), pp. 2494–2498
118. D. Povey, M. Hannemann, G. Boulianne, L. Burget, A. Ghoshal, M. Janda, M. Karafiat, S. Kombrink, P. Motlicek, Y. Qian, K. Riedhammer, K. Vesely, N. T. Vu, in *Proceedings of ICASSP*. Generating exact lattices in the WFST framework (IEEE, USA, 2012), pp. 4213–4216
119. C. Garcia-Mateo, J. Dieguez-Tirado, L. Docio-Fernandez, A. Cardenal-Lopez, in *Proceedings of LREC*. Transcrigal: A bilingual system for automatic indexing of broadcast news (ELRA, Belgium, 2004)
120. A. Moreno, L. Campillos, in *Proceedings of Iberspeech*. MAVIR: a corpus of spontaneous formal speech in spanish and english (ISCA, France, 2004), pp. 224–230
121. A. Stolcke, in *Proceedings of Interspeech*. SRILM - an extensible language modeling toolkit (ISCA, France, 2002), pp. 901–904
122. E. Rodríguez-Banga, C. Garcia-Mateo, F. Méndez-Pazó, M. González-González, C. Magariños, in *Proceedings of Iberspeech*. Cotovía: an open source TTS for galician and spanish (ISCA, France, 2012), pp. 308–315
123. D. Can, M. Saraclar, Lattice indexing for spoken term detection. IEEE Trans. Audio, Speech Lang. Process. **19**(8), 2338–2347 (2011)
124. J. Parapar, A. Freire, A. Barreiro, in *Proceedings of ECIR*. Revisiting n-gram based models for retrieval in degraded large collections (Springer, Germany, 2009), pp. 680–684
125. P. Lopez-Otero, J. Parapar, A. Barreiro, Efficient query-by-example spoken document retrieval combining phone multigram representation and dynamic time warping. Inf. Process. Manag. **56**(1), 43–60 (2019)
126. J. Ponte, W. Croft, in *Proceedings of ACM SIGIR*. A language modeling approach to information retrieval (ACM, USA, 1998), pp. 275–281
127. C. Manning, P. Raghavan, H. Schutze, *Introduction to Information Retrieval*. (Cambridge University, Cambridge, 2008)
128. A. Abad, L. J. Rodríguez-Fuentes, M. Peñagarikano, A. Varona, G. Bordel, in *Proceedings of Interspeech*. On the calibration and fusion of heterogeneous spoken term detection systems (ISCA, France, 2013), pp. 20–24
129. N. Brummer, D. van Leeuwen, in *Proceedings of IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*. On calibration of language recognition scores (IEEE, USA, 2006), pp. 1–8
130. N. Brummer, E. de Villiers, The BOSARIS toolkit user guide: theory, algorithms and code for binary classifier score processing. Agnitio Labs (2011). https://sites.google.com/site/nikobrummer. Accessed 11 Aug 2019
131. D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, A. Rastrow, R. C. Rose, P. Schwarz, S. Thomas, The subspace Gaussian mixture model: a structured model for speech recognition. Comput. Speech Lang. **25**(2), 404–439 (2011)
132. gTTS (Google Text-to-Speech), Python library and CLI tool to interface with Google Translate's text-to-speech API. https://pypi.org/project/gTTS/. Accessed 11 Aug 2019
133. NSSpeechSynthesizer, The cocoa interface to speech synthesis in macOS (AppKit Module of PyObjC Bridge). https://developer.apple.com/documentation/appkit/nsspeechsynthesizer. Accessed 11 Aug 2019
134. Python Interface to the WebRTC (https://webrtc.org/) Voice activity detector (VAD). https://github.com/wiseman/py-webrtcvad. Accessed 11 Aug 2019
135. A. Silnova, P. Matejka, O. Glembek, O. Plchot, O. Novotny, F. Grezl, P. Schwarz, L. Burget, J. H. Cernocky, in *Proceedings of Odyssey*. BUT/Phonexia bottleneck feature extractor (IEEE, USA, 2018), pp. 283–287
136. C. Cieri, D. Miller, K. Walker, in *Proceedings of LREC*. The Fisher Corpus: a resource for the next generations of speech-to-text (ELRA, Belgium, 2004), pp. 69–71
137. Intelligence Advanced Research Projects Activity (IARPA): Babel program. Intelligence Advanced Research Projects Activity (IARPA). https://www.iarpa.gov/index.php/research-programs/babel. Accessed 11 Aug 2019
138. L. J. Rodriguez-Fuentes, A. Varona, M. Penagarikano, G. Bordel, M. Diez, in *Proceedings of ICASSP*. High-performance query-by-example spoken term detection on the SWS 2013 evaluation (IEEE, USA, 2014), pp. 7819–7823
139. A. Abad, L. J. Rodriguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, G. Bordel, in *Proceedings of Interspeech*. On the calibration and fusion of heterogeneous spoken term detection systems (ISCA, France, 2013), pp. 20–24

Publisher's Note