


RESEARCH

Open Access



# Search on speech from spoken queries: the Multi-domain International ALBAYZIN 2018 Query-by-Example Spoken Term Detection Evaluation

Javier Tejedor<sup>1\*</sup> , Doroteo T. Toledano<sup>2</sup>, Paula Lopez-Otero<sup>3†</sup>, Laura Docio-Fernandez<sup>4†</sup>, Mikel Peñagarikano<sup>5†</sup>, Luis Javier Rodriguez-Fuentes<sup>5†</sup> and Antonio Moreno-Sandoval<sup>6</sup>

## Abstract

The huge amount of information stored in audio and video repositories makes search on speech (SoS) a priority area nowadays. Within SoS, Query-by-Example Spoken Term Detection (QbE STD) aims to retrieve data from a speech repository given a spoken query. Research on this area is continuously fostered with the organization of QbE STD evaluations. This paper presents a multi-domain internationally open evaluation for QbE STD in Spanish. The evaluation aims at retrieving the speech files that contain the queries, 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: MAVIR database, which comprises a set of talks from workshops; RTVE database, which includes broadcast television (TV) shows; and COREMAH database, which contains 2-people spontaneous speech conversations about different topics. The evaluation has been designed carefully so that several analyses of the main results can be carried out. We present the evaluation itself, the three databases, the evaluation metrics, the systems submitted to the evaluation, the results, and the detailed post-evaluation analyses based on some query properties (within-vocabulary/out-of-vocabulary queries, single-word/multi-word queries, and native/foreign queries). Fusion results of the primary systems submitted to the evaluation are also presented. Three different teams took part in the evaluation, and ten different systems were submitted. The results suggest that the QbE STD task is still in progress, and the performance of these systems is highly sensitive to changes in the data domain. Nevertheless, QbE STD strategies are able to outperform text-based STD in unseen data domains.

**Keywords:** Query-by-Example Spoken Term Detection, International evaluation, Spanish language, Search on speech

## 1 Introduction

The huge amount of information stored in audio and audiovisual repositories makes it necessary to develop efficient methods for search on speech (SoS). Significant research has been carried out in this area from spoken document retrieval (SDR) [1–6], keyword spotting (KWS) [7–12], spoken term detection (STD) [13–18],

and Query-by-Example Spoken Term Detection (QbE STD) [19–26] tasks.

STD aims to find *terms* within audio archives. It is based on a text-based input, commonly the word/phone transcription of the search term, and hence, STD is also called text-based STD. Query-by-Example Spoken Term Detection also aims to search within audio archives but is based on an acoustic (spoken) input. This is a highly valuable alternative for visually impaired people or when using devices that do not have a text-based input (such as smart speakers), and consequently, the query must be given in another format such as speech.

STD systems typically comprise three different stages: (1) the audio is decoded into word/subword lattices using

\*Correspondence: [javier.tejedor@ceu.es](mailto:javier.tejedor@ceu.es)

<sup>†</sup>Paula Lopez-Otero, Laura Docio-Fernandez, Mikel Peñagarikano and Luis Javier Rodriguez-Fuentes contributed equally to this work.

<sup>1</sup>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

an automatic speech recognition (ASR) subsystem trained for the target language; (2) a term detection subsystem searches the terms within those word/subword lattices to hypothesize detections; and (3) confidence measures are computed to rank detections. The STD systems are normally language-dependent and require large amounts of resources to be built.

On the other hand, QbE STD has been traditionally addressed using three different approaches: methods based on the word/subword transcription of the query, methods based on template matching of features, and hybrid approaches. These approaches are described below.

### 1.1 Methods based on the word/subword transcription of the spoken query

In these methods, first, the spoken query is decoded using an ASR system and then a text-based STD approach is employed to hypothesize detections. The errors produced in the transcription of the query can lead to significant performance degradation. In [21] and [27], the authors employ a Viterbi-based search on hidden Markov models (HMMs). In other works [19, 28–30] dynamic time warping (DTW) or variants of DTW are applied (e.g., non-segmental dynamic time warping (NS-DTW)) to align phone sequences. More sophisticated approaches [20, 31–33] employ word and syllable speech recognizers. In [34], the authors employ a phone-based speech recognizer and weighted finite state transducer (WFST)-based search, whereas in [35], they apply multilingual phone-based speech recognition from supervised and unsupervised acoustic models and sequential dynamic time warping for search. The works [36–38] propose the discovery of unsupervised acoustic features (e.g., bottleneck features) and unsupervised acoustic units for query/utterance representation, and [39] and the work by (Lopez-Otero et al.: Probabilistic information retrieval models for query-by-example spoken document retrieval, submitted to *Multimed. Tools Appl.*) make use of information retrieval models for QbE STD employing ASR.

### 1.2 Methods based on template matching

In these methods, sequences of feature vectors are extracted from both the input spoken queries and the utterances, which are then used in the search stage to hypothesize detections. Regarding the features used for query/utterance representation, Gaussian posteriorgrams are employed in [22, 29, 40, 41]; an *i*-vector-based approach for feature extraction is proposed in [42]; phone log-likelihood ratio-based features are used in [43]; posteriorgrams derived from various unsupervised tokenizers, supervised tokenizers, and semi-supervised tokenizers are employed in [44]; and posteriorgrams derived from a Gaussian mixture model (GMM) tokenizer, phoneme

recognition, and acoustic segment modeling are used in [45]. Phoneme posteriorgrams have been widely used [34, 41, 46–54] and bottleneck features as well [34, 55–60]. Posteriorgrams from non-parametric Bayesian models are used in [61], articulatory class-based posteriorgrams are employed in [62], intrinsic spectral analysis is proposed in [63], unsupervised segment-based bag of acoustic words is employed in [64], and [65] is based on the sparse subspace modeling of posteriorgrams. An exhaustive feature set is proposed in [66], which includes Mel-frequency cepstral coefficients (MFCCs), spectral entropy, fundamental frequency, among others.

All these studies employ the standard DTW algorithm for query search, except for [40], which employs the NS-DTW algorithm, [41, 50, 51, 53, 56, 59, 61, 66] which employ the subsequence DTW (S-DTW) algorithm, [22] which presents a variant of the S-DTW algorithm, and [52] which employs the segmental DTW algorithm. An interesting alternative is [54] which proposes the use of hashing of the phone posteriors to speed-up search and to enable searching on massively large datasets.

These template matching-based methods were found to outperform subword transcription-based techniques in QbE STD [67] and can be effectively employed to build language-independent STD systems, since prior knowledge of the language involved in the speech data is not necessary.

### 1.3 Hybrid methods

These methods take advantage of the text-based STD approach and the approaches based on template matching by combining them to hypothesize detections. A powerful way of enhancing the performance relies on building hybrid (fused) systems that combine the two individual methods. Logistic regression-based fusion of acoustic keyword spotting and DTW-based systems using language-dependent phoneme recognizers is presented in [68–70]. An information retrieval technique to hypothesize detection and DTW-based score detection are proposed in [39]. Logistic regression-based fusion on DTW and phone-based systems is employed in [71–74]. DTW-based search at the HMM state-level from syllables obtained from a word-based speech recognizer and a deep neural network (DNN) posteriorgram-based rescoring are employed in [75], and [76] adds a logistic regression-based approach for detection rescoring. Finally, [77] employs a syllable-based speech recognizer and dynamic programming at the triphone state level to output detections and DNN posteriorgram-based rescoring.

## 2 Methods

Research carried out in a certain area may be difficult to compare in the absence of a common evaluation framework. In QbE STD, research also suffers from this issue

since the published systems typically employ different acoustic databases and different lists of queries that make system comparison impossible. In this context, international evaluations provide a unique framework to measure the progress of any technology, such as QbE STD in this case.

ALBAYZIN evaluation campaigns comprise an internationally open set of evaluations supported by the Spanish Thematic Network on Speech Technologies (RTTH<sup>1</sup>) and the ISCA Special Interest Group on Iberian Languages (SIG-IL<sup>2</sup>), 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 [78–87].

Spanish is a major language in the world, and significant research has been conducted on it for ASR, KWS, and STD tasks [88–94]. The increasing interest in SoS around the world and the lack of SoS evaluations dealing with Spanish encouraged us to organize a series of QbE 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 QbE STD evaluation is the addition of a new data domain, namely broadcast television (TV) shows, with the inclusion of shows 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 data domain. Moreover, the queries used in one of the databases (MAVIR) in the ALBAYZIN 2016 QbE 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 QbE STD multi-domain evaluation whose systems are ranked according to different databases and different domains
- Provide an evaluation and benchmark with increasing complexity in the search queries compared to the previous ALBAYZIN QbE STD evaluations

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

This paper is organized as follows: First, Section 3 presents the evaluation and a comparison with other QbE STD evaluations. Then, Section 4, the different systems submitted to the evaluation, along with a text-based STD system, are presented. Evaluation results and discussion are presented in Section 5, which includes the corresponding paired  $t$  tests [95] as statistical significance measure for system comparison. The Section 6 presents a

post-evaluation analysis based on some properties of the queries and the fusion of the primary systems submitted to the evaluation. The last section outlines the main conclusions of the paper.

### 3 ALBAYZIN 2018 QbE STD evaluation

#### 3.1 Evaluation overview

This evaluation involves searching queries given in spoken form within speech data, by indicating the appropriate audio files with the occurrences and timestamps that contain any of those queries.

The evaluation consists in searching different query lists within different sets of speech data. Speech data comprise different domains (workshop talks, broadcast TV shows, and 2-people conversations), for which individual datasets are given. The ranking of the evaluation results is based on the average system performance on the three datasets in the test experiments.

Two different types of queries are defined in this evaluation, in-vocabulary (INV) and out-of-vocabulary (OOV) queries. The OOV query set was defined to simulate the out-of-vocabulary words of a large vocabulary continuous speech recognition (LVCSR) system. In case participants employ LVCSR for processing the audio, these OOV words must be previously removed from the system dictionary, and hence, other methods have to be used for searching OOV queries. On the other hand, the INV queries could appear in the LVCSR system dictionary.

Participants could submit a primary system and up to four 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 [96].

About 3 months were given to the participants for system development, and therefore, the QbE STD evaluation focuses on building QbE 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 by the end of June 2018. The test data were released by the beginning of September 2018. The final system submission was due by mid-October 2018. Final results were discussed at IberSPEECH 2018 conference by the end of November 2018.

#### 3.2 Evaluation metrics

In QbE 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* (FA). If an actual occurrence is not detected, it is called a *miss*. The actual term-weighted value (ATWV) metric proposed by the National Institute of Standards and Technology (NIST) [96] has been used as the main metric for the evaluation. This metric integrates the hit rate and false alarm rate of each query

into a single metric and then averages over all the queries:

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

where  $\Delta$  denotes the set of queries and  $|\Delta|$  is the number of queries in this set.  $N_{hit}^K$  and  $N_{FA}^K$  represent the numbers of hits and false alarms of query  $K$ , respectively, and  $N_{true}^K$  is the number of actual occurrences of  $K$  in the audio.  $T$  denotes the audio length in seconds, and  $\beta$  is a weight factor set to 999.9 as in [97]. 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 QbE STD system (usually tuned on development data). An additional metric, called maximum term-weighted value (MTWV) [96], can also be used to evaluate the performance of a QbE STD system. MTWV is the maximum TWV obtained by the QbE 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 QbE STD system. Results based on this metric are also presented to evaluate the system performance regardless of the decision threshold.

In addition to ATWV and MTWV, NIST also proposed a detection error tradeoff (DET) curve [98] to evaluate the performance of a QbE 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 [99] was employed to compute MTWV, ATWV, and DET curves.

### 3.3 Databases

Three different databases that comprise different acoustic conditions and domains have been employed for the evaluation: (1) MAVIR database, which was employed in all the previous ALBAYZIN QbE STD evaluations, is used for comparison purposes; (2) RTVE database, which consists of different programs recorded from the Spanish public television (Radio Televisión Española) and involves different broadcast TV shows; (3) COREMAH database, which contains conversational speech with two speakers per recording. For MAVIR and RTVE databases, three separate datasets (i.e., training, development, and test) were provided to the participants. For COREMAH database, only test data were provided. This allowed measuring the generalization capability of the systems in an unseen data domain. Tables 1, 2, and 3 include some database features such as the division into training, development, and test data of the speech files; the number of word occurrences; duration; the number of speakers; and average

**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	#word occ.	dur. (min)	#spk.	Ave. MOS
Mavir-02	train	13432	74.51	7 (7 ma.)	2.69
Mavir-03	dev	6681	38.18	2 (1 ma. 1 fe.)	2.83
Mavir-06	train	4332	29.15	3 (2 ma. 1 fe.)	2.89
Mavir-07	dev	3831	21.78	2 (2 ma.)	3.26
Mavir-08	train	3356	18.90	1 (1 ma.)	3.13
Mavir-09	train	11179	70.05	1 (1 ma.)	2.39
Mavir-12	train	11168	67.66	1 (1 ma.)	2.32
Mavir-04	test	9310	57.36	4 (3 ma. 1 fe.)	2.85
Mavir-11	test	3130	20.33	1 (1 ma.)	2.46
Mavir-13	test	7837	43.61	1 (1 ma.)	2.48
ALL	train	43467	260.27	13 (12 ma. 1 fe.)	2.56
ALL	dev	10512	59.96	4 (3 ma. 1 fe.)	2.64
ALL	test	20277	121.3	6 (5 ma. 1 fe.)	2.65

ma. male, fe. female. These characteristics are displayed for training (train), development (dev), and testing (test) datasets

mean opinion score (MOS) [100] as a way to get an idea of the quality of each speech file in the different databases.

#### 3.3.1 MAVIR

MAVIR database consists of a set of Spanish talks extracted from the MAVIR workshops<sup>3</sup> held in 2006, 2007, and 2008.

The MAVIR Spanish data consist of spontaneous speech files from different speakers from Spain and Latin America, 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, the organizers manually added the timestamps for the roughly 1600 occurrences of the spoken queries 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<sup>4</sup>.

The speech data were originally recorded in several audio formats (pulse-code modulation (PCM) mono and stereo, MP3, 22.05 kHz, 48 kHz, among others). The recordings were converted to PCM, 16 kHz, single channel, 16 bits per sample using the SoX tool<sup>5</sup>. 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 speaker's mouth to the microphone varied and was not controlled at all, but it was smaller

**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	#word occ.	dur. (min)	#spk.	Ave. MOS
LN24H-20151125	dev2	21049	123.50	22	3.37
LN24H-20151201	dev2	19727	112.43	16	3.27
LN24H-20160112	dev2	18617	110.40	19	3.24
LN24H-20160121	dev2	18215	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	3729924	27729	N/A	3.04
ALL	dev1	545952	3742.88	N/A	2.90
ALL	dev2	149255	930.64	N/A	3.25
ALL	test	90021	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 (only for testing). Number of word occurrences (#occ.), duration (dur.) in seconds (sec), 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

ma. male, fe. female

than 50 cm in most of the cases. 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.

### 3.3.2 RTVE

RTVE database belongs to the broadcast TV program domain and contains speech from different TV shows recorded from 2015 to 2018 (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 1400 occurrences of the spoken queries used in the development and test evaluation sets. The training data were available to participants with the corresponding subtitles (note that subtitles are not literal transcriptions of speech data). The development data were further divided into two different development sets: the *dev1* dataset consists of about 60 h of speech material with human-revised word transcriptions without time alignment and the *dev2* dataset, the one that was actually employed as *real* development data for QbE 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 SoS evaluation in Spanish language. More information about the RTVE database can be found in [101].

### 3.3.3 COREMAH

COREMAH database contains conversations about different topics such as rejection, compliment, and apology. It was recorded in 2014 and 2015 in a university environment<sup>6</sup> [102]. This database contains about 45 min of speech data from speakers with different levels of fluency in Spanish (native, intermediate B1, and advanced C1). Since the main purpose of this database is to evaluate the submitted systems in an unseen data domain, only the Spanish native speaker recordings are employed in the evaluation in order to recreate the same conditions of the other databases. To prepare the data for the evaluation, organizers manually added the timestamps for the roughly 850 occurrences of the spoken queries used in the test evaluation set.

The original recordings are videos in 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*<sup>7</sup> tool. It is worth mentioning the large degree of overlapped speech in the recordings, which makes this database very challenging for the evaluation.

### 3.3.4 Query list selection

The selection of queries for the development and test sets aimed to build a realistic scenario for QbE STD by including high-occurrence queries, low-occurrence queries, in-language (INL) (i.e., Spanish) queries, out-of-language (OOL) (i.e., foreign) queries, single-word and multi-word queries, in-vocabulary and out-of-vocabulary queries, and queries of different length. A query may not have any occurrence or appear one or more times in the development/test speech data. Table 4 presents some relevant features of the development and test lists of queries such as the number of INL and OOL queries, the number of single-word and multi-word queries, and the number of INV and OOV queries, along with the number of occurrences of each type in the corresponding dataset. It must be noted that a multi-word query is considered OOV in case any of the words that form the query is OOV.

### 3.4 Comparison to other QbE STD international evaluations

The QbE STD evaluations that are the most similar to ALBAYZIN are MediaEval 2011 [103], 2012 [104], and 2013 [105] Spoken Web Search (SWS) evaluations. However, these evaluations differ in several aspects:

- The most important difference is the nature of the audio content used for the evaluations. In the SWS evaluations, the speech was typically telephone speech, either conversational or read and elicited speech, or speech recorded with in-room microphones. In the ALBAYZIN QbE STD evaluations, the audio consisted of microphone recordings of real talks in workshops that took place in large conference rooms in the presence of an audience. In addition, ALBAYZIN 2018 QbE STD evaluation also contains live-talking conversational speech and broadcast TV shows and explicitly defines different in-vocabulary and out-of-vocabulary query sets.
- SWS evaluations dealt with Indian- and African-derived languages, as well as Albanian, Basque, Czech, non-native English, Romanian, and Slovak languages, while the ALBAYZIN QbE STD evaluations only deal with Spanish language.

These differences make it difficult to compare the results obtained in ALBAYZIN and SWS QbE STD evaluations.

In 2014, the Query-by-Example Search on Speech Task (QUESST) held at MediaEval differed from the previous evaluations in that it was a spoken document retrieval task (i.e., no query timestamps had to be output) [106]. In 2015, QUESST was similar to that of 2014, but the acoustic conditions of the speech data were much more complicated

**Table 4** Characteristics of the lists of development and test queries for MAVIR, RTVE, and COREMAH databases

Query list	dev-MAVIR	dev-RTVE	test-MAVIR	test-RTVE	test-COREMAH
#IN-LANG queries (occ.)	96 (386)	81 (464)	99 (1163)	89 (808)	89 (849)
#OUT-LANG queries (occ.)	6 (39)	22 (110)	7 (29)	19 (72)	2 (8)
#SINGLE queries (occ.)	93 (407)	101 (544)	100 (1180)	105 (861)	90 (840)
#MULTI queries (occ.)	9 (18)	2 (30)	6 (12)	3 (19)	1 (17)
#INV queries (occ.)	83 (296)	76 (480)	94 (979)	87 (750)	81 (800)
#OOV queries (occ.)	19 (129)	27 (94)	12 (213)	21 (130)	10 (57)

dev development, *IN-LANG* in-language queries, *OUT-LANG* foreign queries, *SINGLE* single-word queries, *MULTI* multi-word queries, *INV* in-vocabulary queries, *OOV* out-of-vocabulary queries, *occ.* occurrences

(e.g., reverberation, different kinds of noise), and there were different types of queries (exact queries, queries with lexical variations, queries with changes in the word order, to name a few) [107].

In addition to the MediaEval evaluations, other QbE STD evaluations were organized with the NTCIR-11 [108] and NTCIR-12 [109] conferences. The data used in these evaluations contained spontaneous speech in Japanese provided by the National Institute for Japanese Language, and spontaneous speech recorded during seven editions of the Spoken Document Processing Workshop. As additional information, these evaluations provided the participants with the results of a voice activity detection (VAD) system for the speech data, the manual transcription of the speech data, and the output of an LVCSR system. Although ALBAYZIN QbE STD evaluations are somehow similar in terms of speech nature to the NTCIR QbE STD evaluations (i.e., the speech was recorded in real workshops), ALBAYZIN QbE STD evaluations make use of a different language and define disjoint development and test query lists to measure the generalization capability of the systems.

Table 5 summarizes the main characteristics of SWS, NTCIR, and ALBAYZIN QbE STD evaluations.

**Table 5** Comparison of the different QbE STD evaluations

Evaluation	Language/s	Type of speech	# queries dev./test	Primary metrics
MediaEval 2011	English, Hindi, Gujarati, and Telugu	Tel.	64/36	ATWV
MediaEval 2012	2011 + isiNdebele, Siswati, Tshivenda, and Xitsonga	Tel.	164/136	ATWV
MediaEval 2013	ALB, BAS, CZE, NN-ENG, ISIX, ISIZ, ROM, SEP, and SET	Tel. and mic.	> 600/> 600	ATWV
MediaEval 2014	ALB, BAS, CZE, NN-ENG, ROM, and SLO	Tel. and mic.	560/555	$C_{nxe}$
NTCIR-11 2014	Japanese	mic. workshop	63/203	$F$ -measure
NTCIR-12 2016	Japanese	mic. workshop	120/1620	$F$ -measure ATWV MAP
ALBAYZIN 2012	Spanish	mic. workshop	60/60	ATWV
ALBAYZIN 2014	Spanish	mic. workshop	94/99	ATWV
ALBAYZIN 2016	Spanish	mic. workshop+parliament	102/106+95	ATWV
ALBAYZIN 2018	Spanish	mic. workshop+BNews+conv.	102 + 103/106 + 108 + 91	ATWV

Tel. telephone, mic. microphone, BNews broadcast news, conv. conversational, dev. development, ATWV actual term-weighted value,  $C_{nxe}$  normalized cross entropy cost, MAP mean average precision, ALB Albanian, BAS Basque, CZE Czech, NN-ENG non-native English, ISIX Isixhosa, ISIZ Isizulu, ROM Romanian, SEP Sepedi, SET Setswana, SLO Slovak

## 4 Systems

Three teams submitted ten different systems to ALBAYZIN 2018 QbE STD evaluation, as listed in Table 6. The systems belong to three of the categories described above: text-based STD, template matching, and hybrid systems.

### 4.1 A-Hybrid DTW+LVCSR system

This system (Fig. 1) consists of the fusion of four different QbE STD systems. Three of them are based on DTW, and the other on LVCSR.

#### 4.1.1 Feature extraction in DTW-based systems

Each DTW-based system employs a different speech representation:

- Phoneme posteriorgrams [67], which represent the probability of each phonetic unit at every time instant. The English phone decoder developed by the Brno University of Technology (BUT) [110] is used to obtain phoneme posteriorgrams, and then a Gaussian softening is applied in order to have Gaussian-distributed probabilities [111].
- Low-level descriptors (Table 7) obtained using the OpenSMILE feature extraction toolkit [112]) are

**Table 6** Participants in ALBAYZIN 2018 QbE STD evaluation along with the submitted systems

Team ID	Research institution	Systems	Type of system
GTM-IRLab	AtlantTIC Research Center + Information Retrieval Lab. Universidade de Vigo + Universidade da Coruña, Spain	A-Hybrid DTW+LVCSR	LD hybrid
		B-Fusion DTW	LI template matching
		C-PhonePost DTW	LI template matching
		D-LVCSR	LD LVCSR
AUDIAS-CEU	Universidad Autónoma de Madrid + Universidad CEU San Pablo, Spain	E-DTW	LI template matching
GTTS	Universidad del País Vasco, Spain	F-Combined DTW	LI template matching
		G-Super-BNF DTW	LI template matching
		H-Multilingual-BNF DTW	LI template matching
		I-Monoph.-BNF DTW	LI template matching
		J-Triph.-BNF DTW	LI template matching

LD language-dependent, LI language-independent

extracted every 10 ms using a 25-ms window, except for F0, probability of voicing, jitter, shimmer, and harmonics-to-noise ratio (HNR), for which a 60-ms window is used. These features are augmented with their delta coefficients.

- Gaussian posteriorgrams [113], which represent the probability of each Gaussian in a GMM at every time instant. Feature extraction and Gaussian posteriorgram computation are performed using the Kaldi toolkit [114]. The GMM is trained employing MAVIR and RTVE training as well as RTVE *dev1* data, using 19 MFCCs plus energy, and their delta and delta delta coefficients.

#### 4.1.2 Query detection

From each feature set described above, a search procedure is followed to hypothesize query detections. The search is based on the S-DTW algorithm [115], which is a variant

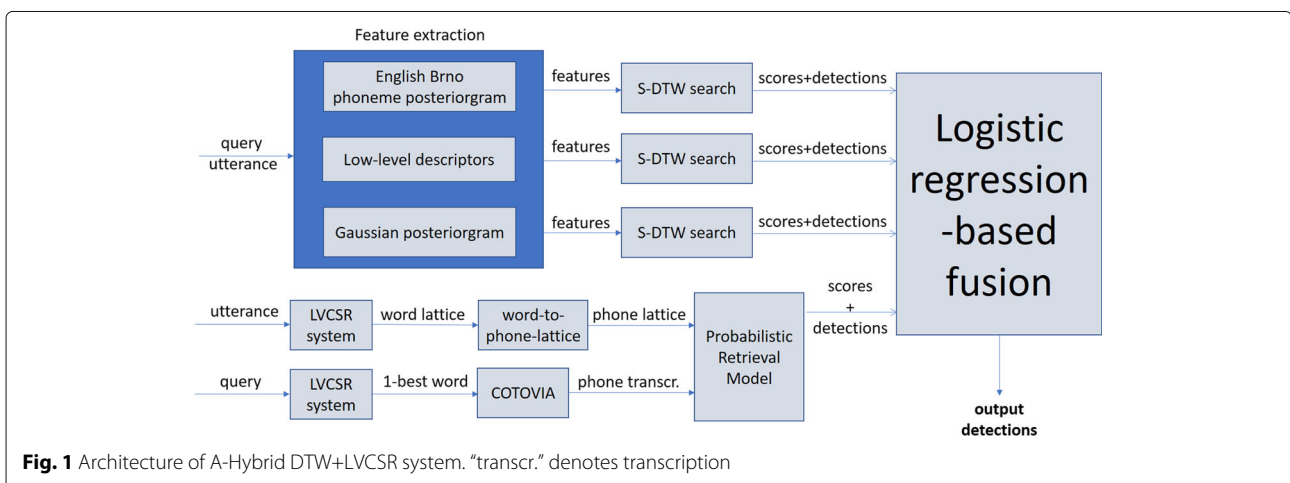
of the standard DTW search. In S-DTW, a cost matrix  $M \in \mathbb{R}^{n \times m}$  must first be defined, in which the rows and the columns correspond to the frames of the query ( $Q$ ) and the utterance ( $U$ ), respectively:

$$M_{ij} = \begin{cases} c(q_i, u_j) & \text{if } i = 0 \\ c(q_i, u_j) + M_{i-1,0} & \text{if } i > 0, j = 0 \\ c(q_i, u_j) + M^*(i, j) & \text{otherwise,} \end{cases} \quad (2)$$

where  $c(q_i, u_j)$  is a function that defines the cost between the query vector  $q_i$  and the utterance vector  $u_j$ , and

$$M^*(i, j) = \min(M_{i-1, j}, M_{i-1, j-1}, M_{i, j-1}), \quad (3)$$

which implies that only horizontal, vertical, and diagonal path movements are allowed.



**Fig. 1** Architecture of A-Hybrid DTW+LVCSR system. "transcr." denotes transcription



**Table 7** Acoustic features used in the *A-Hybrid DTW+LVCSR QbE STD* system

Description	Number of features
Sum of auditory spectra	1
Zero-crossing rate	1
Sum of RASTA style filtering auditory spectra	1
Frame intensity	1
Frame loudness	1
Root mean square energy and log-energy	2
Energy in frequency bands 250–650 Hz (energy 250–650) and 1000–4000 Hz	2
Spectral Rolloff points at 25%, 50%, 75%, 90%	4
Spectral flux	1
Spectral entropy	1
Spectral variance	1
Spectral skewness	1
Spectral kurtosis	1
Psychoacoustical sharpness	1
Spectral harmonicity	1
Spectral flatness	1
Mel-frequency cepstral coefficients	16
MFCC filterbank	26
Line spectral pairs	8
Cepstral perceptual linear predictive coefficients	9
RASTA PLP coefficients	9
Fundamental frequency (F0)	1
Probability of voicing	1
Jitter	2
Shimmer	1
log harmonics-to-noise ratio (logHNR)	1
LCP formant frequencies and bandwidths	6
Formant frame intensity	1
Deltas	102
Total	204

*PLP* perceptual linear predictive, *LPC* linear predictive coding

Pearson's correlation coefficient  $r$  [116] is used as a cost function by mapping it into the interval  $[0,1]$  applying the following transformation:

$$c(q_i, u_j) = \frac{1 - r(q_i, u_j)}{2}. \quad (4)$$

Once the matrix  $M$  is computed, the end of the best warping path between  $Q$  and  $U$  is obtained as follows:

$$b^* = \arg \min_{b \in \{1, \dots, m\}} M(n, b). \quad (5)$$

The starting point of the path ending at  $b^*$ , namely  $a^*$ , is computed by backtracking, hence obtaining the best warping path.

A query  $Q$  may appear several times in an utterance  $U$ , especially if  $U$  is a long recording. Therefore, not only must the best warping path be detected, but also others that are less likely. One approach to overcome this issue consists in detecting a given number of candidate matches  $n_c$ : Every time a warping path that ends at frame  $b^*$  is detected,  $M(n, b^*)$  is set to  $\infty$  to ignore this element in the future.

A confidence score must be assigned to every detection of a query  $Q$  in an utterance  $U$ . Firstly, the cumulative cost of the warping path  $M_{n,b^*}$  is length-normalized [68] and then  $z$ -norm is applied so that all the confidence scores of all the queries have the same distribution [70].

#### 4.1.3 LVCSR-based QbE STD

This strategy follows a probabilistic retrieval model for information retrieval [117] that is applied in this evaluation for the QbE STD task. This model consists of the following stages:

- Indexing: A DNN-based LVCSR system built with the Kaldi toolkit [114] is employed. The utterances are converted into phone-level  $n$ -best lists to store different phone transcriptions (50) for each utterance. Then, these are indexed in terms of phone  $n$ -grams of different size [39, 118]. The minimum and maximum sizes of the  $n$ -grams are set to 1 and 5, respectively, according to [39]. With respect to the probabilistic retrieval model, each utterance is represented by means of a language model (LM) [117]. The start time and duration of each phone are also stored in the index.
- Search: The DNN-based LVCSR system is employed to obtain the word transcription of each query. Then, it is converted to phone transcription using the dictionary created with Cotovia software [119] and searched within the different indices. Finally, a score for each utterance is computed following the query likelihood retrieval model [120]. It must be noted that this model sorts the utterances according to how likely it is they contain the query, but the start and end times of the match are required in this task. To obtain these times, the phone transcription of query  $Q$  is aligned to that of utterance  $U$  by computing the minimum edit distance (MED)  $MED(Q, U)$ . 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 to *Multimed. Tools Appl.*):

$$\text{score}(Q, U) = \text{score}_{LM}(Q, U) \cdot \text{score}_{MED}(Q, U), \quad (6)$$

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

$$\text{score}_{MED}(Q, U) = \frac{n_Q - MED(Q, U)}{K}, \quad (7)$$

where  $n_Q$  is the number of phonemes of the query, and  $K$  is the length of the best alignment path.

Indexing and search were performed using Lucene.<sup>8</sup>

#### 4.1.4 Calibration and fusion

Discriminative calibration and fusion [121] are applied in order to combine the outputs of the three DTW systems and that of the LVCSR system. The global minimum score produced by the system for all the queries is used to hypothesize the missing scores. After normalization, calibration and fusion parameters are estimated by logistic regression on the development datasets to obtain improved discriminative and well-calibrated scores [122]. Calibration and fusion training are performed using Bosaris toolkit [123].

The decision threshold, weight of the LM in the DNN-based LVCSR system, and number of  $n$ -best lists in the LVCSR-based QbE STD system are tuned from the combined ground truth labels of the MAVIR and RTVE development data. The rest of the parameters are set based on preliminary experiments.

#### 4.2 B-Fusion DTW system

This system combines the DTW-based systems presented in *A-Hybrid DTW+LVCSR* system.

#### 4.3 C-Phoneme-posteriorgram DTW system (C-PhonePost DTW)

This system only employs DTW search on the phoneme posteriorgrams presented in the *A-Hybrid DTW+LVCSR* system, and hence does not make use of the calibration and fusion stage.

#### 4.4 D-LVCSR system

This system only employs the LVCSR approach described in the *A-Hybrid DTW+LVCSR* system to hypothesize query detections.

#### 4.5 E-DTW system

This system (Fig. 2) integrates two different stages: feature extraction and query detection, which are explained next.

##### 4.5.1 Feature extraction

The English phoneme recognizer developed by BUT [110] is employed to compute phoneme posteriorgrams that represent both the queries and the utterances and is very similar to the posteriorgram features of the former systems, except for the Gaussian softening stage.

##### 4.5.2 Query detection

First, a cost matrix that stores the similarity between every query/utterance pair is computed. The Pearson correlation coefficient  $r(q_n, u_m)$  [116] is employed to build the cost matrix, where  $q_n$  represents the query phoneme posteriorgram frames and  $u_m$  represents the utterance phoneme posteriorgram frames.

The final cost used in the search stage is modified as follows:  $c(q_n, u_m) = 1 - \max(0, r(q_n, u_m))$ . Therefore, for all the Pearson correlation coefficient values lower or equal to 0, the cost will be maximum. The S-DTW algorithm explained in the *A-Hybrid DTW+LVCSR* system is employed to hypothesize detections from this cost matrix. Finally, a neighborhood search is carried out so that all the paths (i.e., query detections) which overlap more than 500 ms from a previously obtained optimal path are rejected in the final system output.

Parameter tuning is carried out using MAVIR development data and then applied to the other datasets.

#### 4.6 F-Combined DTW system

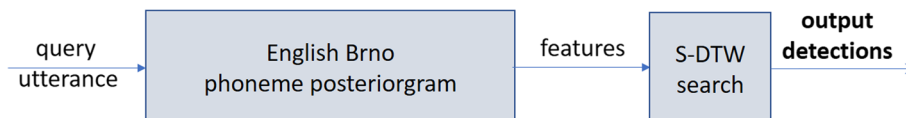
This system (Fig. 3) is based on the combination of different search processes, each employing a different feature set.

##### 4.6.1 Voice activity detection

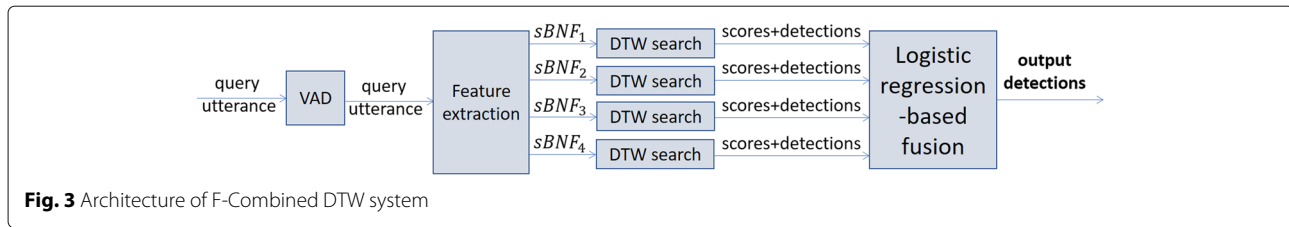
The spoken queries and the utterances are first processed with the VAD system developed by Google for the WebRTC project [124], which is based on Gaussian distributions of speech and non-speech features.

##### 4.6.2 Feature extraction

The feature extraction module performs stacked bottleneck feature (sBNF) computation following the BUT/Phonexia approach [125], both for queries and utterances. To do so, three different neural networks are



**Fig. 2** Architecture of E-DTW system



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 [126] with 120 monophone state targets, referred to as FisherMono. The second one is also trained on the Fisher corpus but with 2423 triphone tied-state targets and is referred to as FisherTri. The third network is trained on telephone speech in 17 languages taken from the Intelligence Advanced Research Projects Activity (IARPA) Babel program [127], with 3096 stacked monophone state targets for the 17 languages involved (BabelMulti for short). Given that the sBNF extractors are trained using 8 kHz speech signals, the queries and the utterances are downsampled to 8 kHz.

The architecture of the sBNF networks consists of two stages. The first one is a standard bottleneck network fed with low-level acoustic features spanning 10 frames (100 ms), the bottleneck size being 80. The second stage takes as input five equally spaced bottleneck features of the first stage, spanning 31 frames (310 ms), and is trained on the same targets as the first stage, with the same bottleneck 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, instead of sBNFs, the extractor can output target posteriors.

The operation of BUT/Phonexia sBNF extractors requires an external VAD module 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.

#### 4.6.3 Dynamic time warping-based search

This system follows the DTW-based approach presented in [128]. Given the two sequences of sBNFs corresponding to a query and an utterance, a VAD system is used to discard non-speech frames, but keeping the timestamp of each frame. To avoid memory issues, utterances 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  $Q = (q[1], q[2], \dots, q[m])$  be the sequence of VAD-filtered sBNFs of length  $m$  corresponding to a query and  $U = (u[1], u[2], \dots, u[n])$  be those of an utterance of length  $n$ . Since sBNFs (theoretically) range from  $-\infty$  to  $+\infty$ , the distance between a pair of vectors  $q[i]$  and  $u[j]$  is defined as follows:

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

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 utterance  $U$  as follows:

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

where

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

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

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 utterance  $U$ .

Note that the chunking process described above makes the normalization procedure differ from that applied in [128], since  $d_{\min}(i)$  and  $d_{\max}(i)$  are not computed for the whole utterance 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 utterance  $U$  of length  $n$  is defined as that minimizing the average distance in a *crossing path* of the matrix  $d_{\text{norm}}$ . A crossing path starts at any given frame of  $U$ ,  $k_1 \in [1, n]$ , then

traverses a region of  $U$  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:

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

where  $i_l$  and  $j_l$  are the indices of the vectors of  $q$  and  $u$  in the alignment  $l$ , for  $l = 1, 2, \dots, L$ . 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, where  $d$  is the size of the feature vectors and  $O(n \cdot m)$  in space. Readers are referred to [128] for more details.

The detection score is computed as  $1 - d_{\text{avg}}(Q, U)$ , 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 utterance.

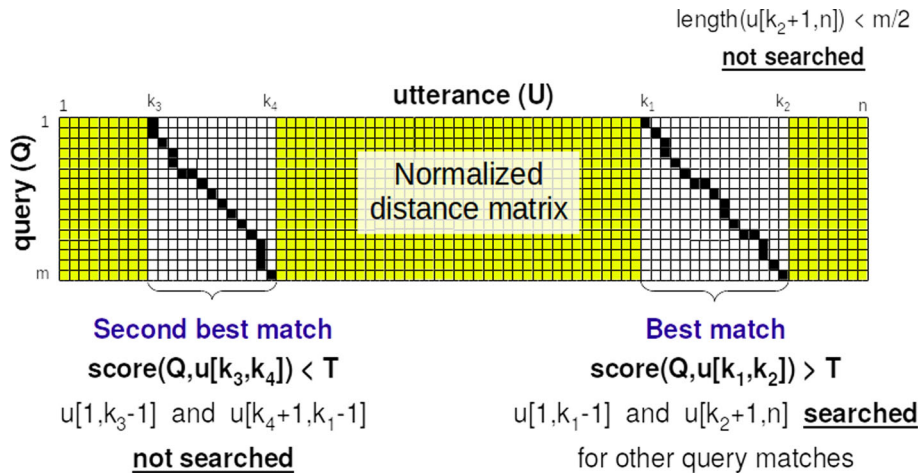
This procedure is iteratively applied to find not only the best match, but also less likely matches in the same utterance. To that end, a queue of search intervals is defined and initialized with  $[1, n]$ . Given an interval  $[a, b]$ , and assuming that the best match is found at  $[a', b']$ , the 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 plus those waiting in the queue) does not exceed a given threshold  $M$  ( $M = 7$ ). An example is shown in Fig. 4. 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, denoted as  $sBNF_1$  in Fig. 3; FisherTri, denoted as  $sBNF_2$  in Fig. 3; and BabelMulti, denoted as  $sBNF_3$  in Fig. 3). The other DTW search employs the concatenation of the three sBNF sets (denoted as  $sBNF_4$  in Fig. 3), which leads to 240-dimensional sBNF vectors. Each DTW search produces different query detections that are next fused in the fusion stage.

#### 4.6.4 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 [129].

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 one trial per second is chosen (which is consistent with the evaluation script provided by the organizers). Given a detection of a query  $Q$  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



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

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.

The calibration/fusion model is estimated on the development set and then applied to both the development and test sets using Bosaris toolkit [123].

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

#### 4.7 G-Super bottleneck feature DTW system (G-Super-BNF DTW)

This system is the same as the *F-Combined DTW* system, except that only DTW-based search on the concatenation of the three sBNFs as features is used to hypothesize query detections.

#### 4.8 H-Multilingual bottleneck feature DTW system (H-Multilingual-BNF DTW)

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

#### 4.9 I-Monophone bottleneck feature DTW system (I-Monoph.-BNF DTW)

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

#### 4.10 J-Triphone bottleneck feature DTW system (J-Triph.-BNF DTW)

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

#### 4.11 K-Text STD system

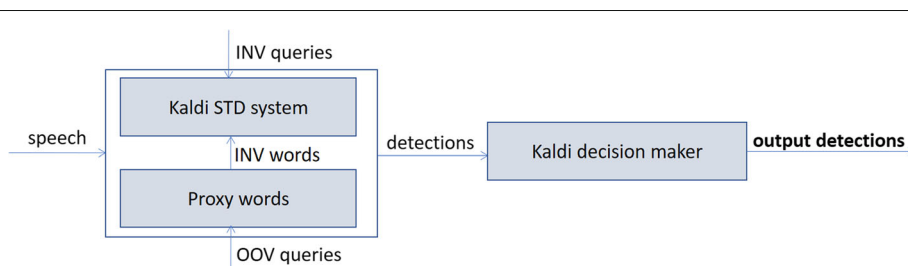
This system (Fig. 5) does not compete in the evaluation itself, but it is presented in order to examine the upper bound limits of QbE STD technology. This system employs the correct word transcription of the query to hypothesize detections and the same ASR approach as that used in the *LVCSR-based QbE STD* system.

The ASR subsystem is based on the Kaldi open-source toolkit [114] and employs DNN-based acoustic models.

The data used to train the acoustic models of this Kaldi-based LVCSR system are extracted from the Spanish material used in the 2006 TC-STAR automatic speech recognition evaluation campaign<sup>9</sup> and the Galician broadcast news database Transcrigal [130]. 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 LM 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, online courses, and the transcriptions of the MAVIR sessions included in the development set provided by the evaluation organizers<sup>10</sup> [131]). Four-gram LMs have been built with the SRILM toolkit [132]. The final LM is an interpolation between a LM trained on RTVE data and another one trained on the rest of the text corpora. The LM vocabulary size is limited to the most frequent 300K words, and for each evaluation dataset, the OOV words are removed from the LM. Grapheme-to-phoneme conversion is carried out with the Cotovia software [119].

The STD subsystem integrates the Kaldi term detector [114, 133, 134], which searches for the input terms within the word lattices obtained in the previous step [135]. The Kaldi decision-maker conducts a YES/NO decision for each detection based on the term-specific threshold approach presented in [136]. To do so, the score for each detection is computed as follows:



**Fig. 5** Architecture of K-Text STD system

$$p > \frac{N_{\text{true}}}{\frac{T}{\beta} + \frac{\beta-1}{\beta} N_{\text{true}}}, \quad (13)$$

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

The proxy words strategy in the Kaldi open-source toolkit [137] is employed for OOV query detection. This strategy consists in substituting each OOV word of the search query with acoustically similar INV proxy words so that the OOV query search can be then carried out using the obtained INV query.

The decision threshold and the weight of the LM in the ASR subsystem 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.

## 5 Evaluation results and discussion

This section presents the overall evaluation results and the results obtained for each individual database on development and test data.

### 5.1 Overall results

The overall evaluation results are presented in Table 8 for development and test data, along with a comparison with the text STD system presented above. These show that the best performance for the ATWV metric on test data is for the *A-Hybrid DTW+LVCSR* system, which highlights the power of hybrid systems for QbE STD. However, two findings arise: (1) the ranking of the evaluation results for development and test data differs and (2) the *K-Text STD* system, which relies on a DNN-based ASR subsystem and the correct word transcription of the query, obtains better results than any of the QbE STD systems in development data, but its performance is similar to that of the best QbE STD system on test data. Calibration threshold issues may be causing these differences in performance, since the *K-Text STD* system also obtains the best MTWV in test data.

### 5.2 Development data

#### 5.2.1 MAVIR

Evaluation results for MAVIR development data are presented in Table 9. By comparing the QbE STD systems, the best performance is obtained with the *B-Fusion DTW* system. Paired  $t$  tests show that the difference in performance is statistically significant ( $p < 0.01$ ) compared with all the QbE STD systems except for the *A-Hybrid*

**Table 8** Overall system results of the ALBAYZIN 2018 QbE STD evaluation on development and test data (average of the results on the two development and three test corpora)

System ID	Development		Test	
	MTWV	ATWV	MTWV	ATWV
A-Hybrid DTW+LVCSR	0.5085	0.4717	0.3260	0.2084
B-Fusion DTW	0.4955	0.4863	0.3082	− 0.4185
C-PhonePost DTW	0.4558	0.4412	0.2891	− 0.4434
D-LVCSR	0.2962	0.2779	0.2080	0.1172
E-DTW	0.1094	0.1070	0.0995	0.0948
F-Combined DTW	0.3758	0.3406	0.1657	0.1413
G-Super-BNF DTW	0.3776	0.3668	0.1460	− 0.0489
H-Multilingual-BNF DTW	0.3379	0.3266	0.1570	− 0.2018
I-Monoph.-BNF DTW	0.3621	0.3342	0.1689	0.0422
J-Triph.-BNF DTW	0.3527	0.3382	0.1628	− 0.0264
K-Text STD	0.6817	0.6480	0.4427	0.2012

*DTW+LVCSR* and *D-LVCSR* systems. *B-Fusion DTW* employs a fusion of DTW-based systems with different feature sets, which suggests that different features convey different patterns that enhance the performance. The *A-Hybrid DTW+LVCSR* system, which integrates an LVCSR-based system in the fusion, does not outperform the *B-Fusion DTW* system, probably due to some threshold calibration issues (better MTWV and worse ATWV) in medium-quality and highly spontaneous speech domains as MAVIR. The *K-Text STD* system, which employs the correct word transcription of the query and an LVCSR approach, performs the best. This best performance is statistically significant for a paired  $t$  test ( $p < 0.01$ ) compared with the rest of the systems. This is due to the use of the correct transcription of the spoken query in the DNN-based speech recognition system, which plays an

**Table 9** System results of the ALBAYZIN 2018 QbE STD evaluation on MAVIR development data

System ID	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$
A-Hybrid DTW+LVCSR	0.2896	0.2470	0.00010	0.606
B-Fusion DTW	0.2699	0.2649	0.00003	0.701
C-PhonePost DTW	0.1971	0.1742	0.00000	0.799
D-LVCSR	0.2383	0.2029	0.00005	0.716
E-DTW	0.1823	0.1774	0.00002	0.801
F-Combined DTW	0.1512	0.0938	0.00005	0.803
G-Super-BNF DTW	0.1629	0.1524	0.00005	0.804
H-Multilingual-BNF DTW	0.1299	0.1076	0.00004	0.828
I-Monoph.-BNF DTW	0.1566	0.1173	0.00002	0.823
J-Triph.-BNF DTW	0.1411	0.1234	0.00001	0.846
K-Text STD	0.6544	0.6042	0.00012	0.229

important role in query detection for highly spontaneous and medium-quality speech domains.

On the other hand, the worst systems are those that employ stacked bottleneck features, which suggests that the use of the sBNFs, as proposed by the authors of those systems, is less powerful than other features for QbE STD in medium-quality and highly spontaneous speech domains.

### 5.2.2 RTVE

The evaluation results for RTVE development data are presented in Table 10. They show that the best performance among the QbE STD systems is obtained with the *C-PhonePost DTW* system. A paired  $t$  test shows that the difference in performance is statistically significant ( $p < 0.01$ ) compared with all the QbE STD systems except for the *A-Hybrid DTW+LVCSR* and *B-Fusion DTW*. This best performance does not correspond to the best system on MAVIR development data, maybe due to the RTVE data comprising higher-quality and better-pronounced speech data than MAVIR, and hence, the best performance may not correspond to the same system. Two more remarked differences can be seen on these data compared to the MAVIR development data: (1) the systems that rely on sBNFs obtain much better performance, and (2) the *K-Text STD* system obtains similar results to that obtained with the *A-Hybrid DTW+LVCSR*, *B-Fusion DTW*, and *C-PhonePost DTW* systems. These differences highlight the power of sBNFs and QbE STD systems when addressing query detection in high-quality and well-pronounced speech domains as RTVE.

Due to threshold calibration issues, the *A-Hybrid DTW+LVCSR* system, which obtains the best MTWV, does not perform the best for ATWV, as in MAVIR development data.

**Table 10** System results of the ALBAYZIN 2018 QbE STD evaluation on RTVE development data

System ID	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$
A-Hybrid DTW+LVCSR	0.7273	0.6964	0.00002	0.254
B-Fusion DTW	0.7211	0.7076	0.00001	0.270
C-PhonePost DTW	0.7145	0.7081	0.00003	0.258
D-LVCSR	0.3540	0.3528	0.00008	0.565
E-DTW	0.0365	0.0365	0.00000	0.963
F-Combined DTW	0.6003	0.5874	0.00006	0.339
G-Super-BNF DTW	0.5922	0.5812	0.00005	0.357
H-Multilingual-BNF DTW	0.5458	0.5455	0.00007	0.387
I-Monoph.-BNF DTW	0.5676	0.5511	0.00008	0.351
J-Triph.-BNF DTW	0.5642	0.5530	0.00004	0.397
K-Text STD	0.7591	0.6918	0.00006	0.184

The *E-DTW* system obtains the worst overall performance. This can be due to the fact that the optimal parameters obtained with the MAVIR development data have been applied on these data without adjustment. Since RTVE data convey many different properties (i.e., high-quality and well-pronounced speech), the parameter tuning is not effective across changes in the data domain.

## 5.3 Test data

### 5.3.1 MAVIR

The results corresponding to MAVIR test data are presented in Table 11. They show that the best performance for QbE STD is obtained with the *B-Fusion DTW* system, which is consistent with the results in MAVIR development data. This best performance is statistically significant for a paired  $t$  test ( $p < 0.01$ ) compared to all the QbE STD systems except for the *C-PhonePost DTW* system, for which the difference is weakly significant ( $p < 0.04$ ). The performance gap between MTWV and ATWV for the best system suggests that the threshold has been well-calibrated. The rest of the findings observed from the development results also arise: (1) the worst systems are those that employ the sBNFs for feature extraction; (2) the *A-Hybrid DTW+LVCSR* system, which integrates an LVCSR approach in the fusion of the *B-Fusion DTW* system, obtains worse performance than the *B-Fusion DTW* system, due to the low performance of the LVCSR system. This indicates that parameter tuning on the development data is not generalizing well on test data; (3) the *K-Text STD* system performs better than any QbE STD system, with all the performance gaps statistically significant for a paired  $t$  test ( $p < 0.01$ ).

### 5.3.2 RTVE

Evaluation results for RTVE test data are presented in Table 12. They show that the best performance for QbE

**Table 11** System results of the ALBAYZIN 2018 QbE STD evaluation on MAVIR test data

System ID	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$
A-Hybrid DTW+LVCSR	0.2226	0.1243	0.00017	0.606
B-Fusion DTW	0.2851	0.2810	0.00007	0.649
C-PhonePost DTW	0.2436	0.2361	0.00003	0.730
D-LVCSR	0.1508	0.1101	0.00008	0.774
E-DTW	0.1550	0.1550	0.00001	0.840
F-Combined DTW	0.1227	0.1157	0.00002	0.860
G-Super-BNF DTW	0.1055	0.0296	0.00002	0.875
H-Multilingual-BNF DTW	0.1277	-0.0026	0.00006	0.814
I-Monoph.-BNF DTW	0.1289	0.0809	0.00002	0.846
J-Triph.-BNF DTW	0.1196	0.0626	0.00004	0.838
K-Text STD	0.5345	0.5178	0.00010	0.364

STD corresponds to the *A-Hybrid DTW+LVCSR* system. The performance gap between MTWV and ATWV on this system indicates that the threshold presents some calibration issues. The best performance is statistically significant for a paired  $t$  test ( $p < 0.01$ ) compared to the rest of the QbE STD systems. This highlights the power of the hybrid systems for QbE STD systems on high-quality and well-pronounced speech data, and for which considerable amount of resources are available. For development data, *A-Hybrid DTW+LVCSR*, *B-Fusion DTW*, and *C-PhonePost DTW* systems obtain equivalent performance. Nevertheless, when test data are given to the hybrid systems, these are able to generalize better than the other systems, due to the complementary information integrated by hybrid systems.

As in development data, the *E-DTW* system performs the worst, due to the fact that no additional tuning on RTVE data has been carried out, whereas the systems that employ sBNFs for feature extraction enhance their performance with respect to the MAVIR test data.

The *K-Text STD* system performs better than any other QbE STD system. This best performance is statistically significant ( $p < 0.01$ ) compared to all the QbE STD systems.

### 5.3.3 COREMAH

Evaluation results for COREMAH test data are presented in Table 13. For the QbE STD systems, the best performance is obtained with the *E-DTW* system. This best performance is statistically significant for a paired  $t$  test ( $p < 0.01$ ) compared to the rest of the QbE STD systems. Remind that no development data were provided for COREMAH, and hence, parameter tuning must be carried out with some other data. The *E-DTW* system was tuned with the MAVIR optimal parameters, which indicates that MAVIR data convey properties which

**Table 13** System results of the ALBAYZIN 2018 QbE STD evaluation on COREMAH test data

System ID	MTWV	ATWV	p(FA)	p(Miss)
A-Hybrid DTW+LVCSR	0.1354	-0.0689	0.00006	0.804
B-Fusion DTW	0.2022	-1.9221	0.00000	0.798
C-PhonePost DTW	0.2059	-1.9062	0.00000	0.794
D-LVCSR	0.0037	-0.1559	0.00000	0.996
E-DTW	0.1436	0.1436	0.00003	0.828
F-Combined DTW	0.0521	0.0023	0.00000	0.948
G-Super-BNF DTW	0.0295	-0.4763	0.00000	0.970
H-Multilingual-BNF DTW	0.0514	-0.8895	0.00000	0.949
I-Monoph.-BNF DTW	0.0538	-0.2779	0.00002	0.930
J-Triph.-BNF DTW	0.0572	-0.4379	0.00000	0.939
K-Text STD	0.0966	-0.5828	0.00007	0.835

are similar to the conversational speech in COREMAH data. However, the parameters of the rest of the systems employed RTVE data for tuning (except for the *F-Combined DTW*, *G-Super-BNF DTW*, *H-Multilingual-BNF DTW*, *I-Monoph.-BNF DTW*, and *J-Triph.-BNF DTW* systems, which employed MAVIR data as well), which leads to a worse performance due to higher data mismatch. For those systems, the low performance may be due to the type of tuning carried out.

The *K-Text STD* system obtains worse performance than the best QbE STD system, although the performance gap is weakly significant for a paired  $t$  test ( $p < 0.03$ ). This could be due to the data mismatch between COREMAH data and the RTVE data, which were used, along with MAVIR data, for parameter tuning in this case.

### 5.4 Analysis of development and test data DET curves

DET curves of the QbE STD systems submitted to the evaluation and the text-based STD system are presented in Figs. 6 and 7 for MAVIR and RTVE development data, respectively, and Figs. 8, 9, and 10 for MAVIR, RTVE, and COREMAH test data, respectively.

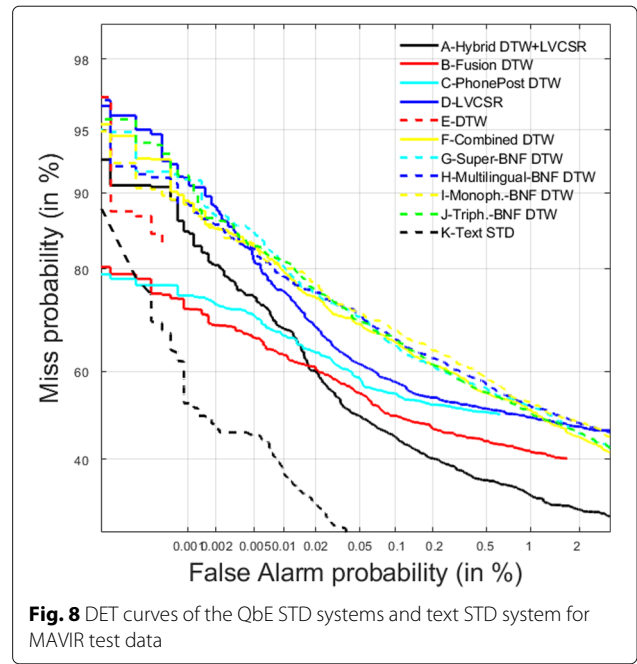
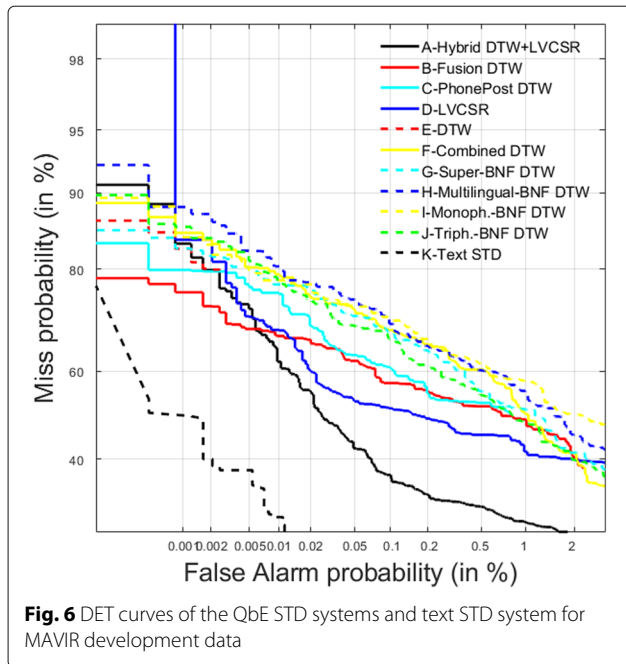
On MAVIR development data, the *B-Fusion DTW* system performs the best for low FA rates, and the *A-Hybrid DTW+LVCSR* system performs the best for moderate and low miss rates. This indicates that the hybrid system is suitable for cases in which a miss is more important than an FA. On RTVE development data, the *B-Fusion DTW* system performs the best for low and moderate FA rates, and the *A-Hybrid DTW+LVCSR* system performs the best for low miss rates. This confirms the power of hybrid systems for low miss rate scenarios.

On MAVIR test data, the *C-PhonePost DTW* system performs the best for very low FA rates, the *B-Fusion DTW* system performs the best for low and moderate

**Table 12** System results of the ALBAYZIN 2018 QbE STD evaluation on RTVE test data

System ID	MTWV	ATWV	p(FA)	p(Miss)
A-Hybrid DTW+LVCSR	0.6201	0.5699	0.00008	0.301
B-Fusion DTW	0.4372	0.3856	0.00004	0.524
C-PhonePost DTW	0.4178	0.3399	0.00002	0.564
D-LVCSR	0.4695	0.3975	0.00008	0.447
E-DTW	0.0000	-0.0141	0.00000	1.000
F-Combined DTW	0.3224	0.3059	0.00003	0.648
G-Super-BNF DTW	0.3029	0.3000	0.00004	0.655
H-Multilingual-BNF DTW	0.2919	0.2868	0.00005	0.662
I-Monoph.-BNF DTW	0.3239	0.3237	0.00005	0.625
J-Triph.-BNF DTW	0.3117	0.2960	0.00003	0.653
K-Text STD	0.6969	0.6685	0.00008	0.226

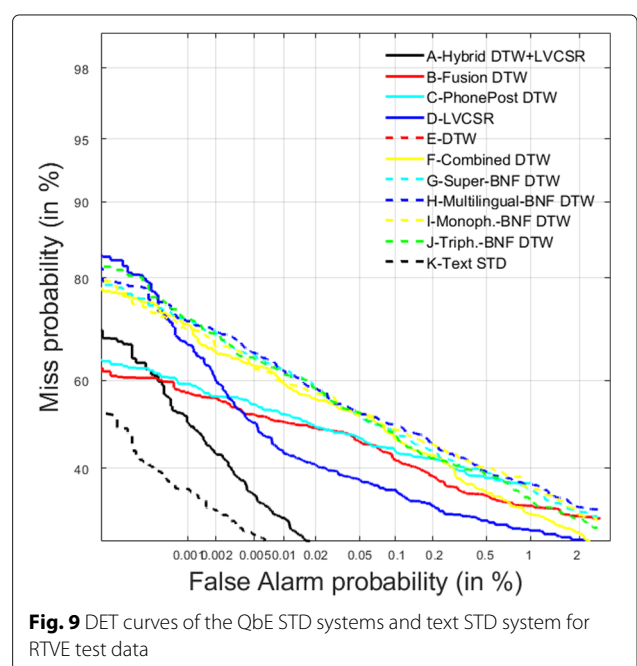
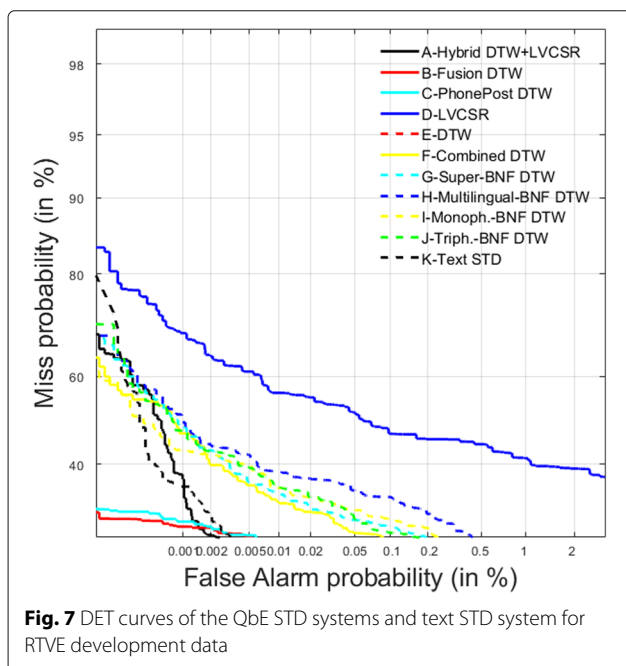


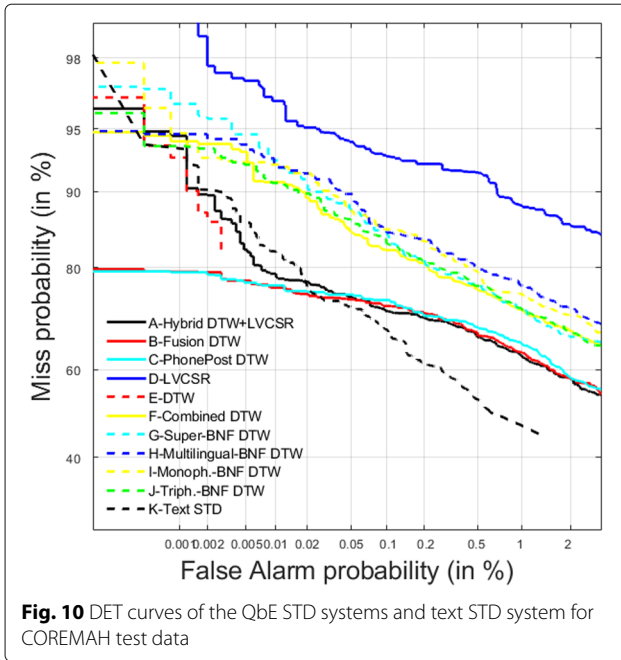


FA rates, and the *A-Hybrid DTW+LVCSR* system performs the best for low miss rates. On RTVE test data, the *B-Fusion DTW* system performs the best for low FA rates, and the *A-Hybrid DTW+LVCSR* system performs the best for low miss rates. On COREMAH test data, the *C-PhonePost DTW* system performs the best for low FA rates, the *B-Fusion DTW* system performs the best for moderate FA rates, and the *A-Hybrid DTW+LVCSR* performs the best for low miss rates. However, according to

the results in Table 13, the best performance is obtained with the *E-DTW* system, since it outputs less number of FAs (although the number of hits is also lower) than those three other systems so that more hits are ranked in top positions, enhancing the MTWV/ATWV performance measure.

In summary, *B-Fusion DTW* and *A-Hybrid DTW+LVCSR* systems obtain the best figures in the DET curves





and make them more appropriate for search on speech from spoken queries.

The DET curves also show that the *K-Text STD* system performs the best for all data, except for RTVE development data on all the operating points, and MAVIR and COREMAH test data on low FA rates. Results on COREMAH test data suggest that QbE STD may outperform text-based STD on unseen data domains, at least on some scenarios (as low FA rates in this case).

## 6 Post-evaluation analysis

After the evaluation period, an analysis based on some query properties and fusion of the primary systems

submitted from the different participants has been carried out. This section presents the results of this analysis.

### 6.1 Performance analysis of QbE STD systems for in-language and out-of-language queries

An analysis of the QbE STD systems and the *K-Text STD* system for in-language and out-of-language queries has been carried out, and the results are presented in Tables 14, 15, and 16 for MAVIR, RTVE, and COREMAH databases, respectively. On MAVIR data, QbE STD system performance is, in general, better on OOL than on INL queries. We consider this is due to the fact all the systems that employ template matching techniques rely on a language-independent approach, for which English language is largely used for feature extraction. Since the OOL queries are in English, this is clearly giving better performance. The systems that employ template matching approaches and obtain better performance on INL than on OOL queries are the *F-Combined DTW*, *G-Super-BNF DTW*, *I-Monoph.-BNF DTW*, and *J-Triph.-BNF DTW* systems, for which the better MTWV performance on OOL queries indicates some threshold calibration issues. The *D-LVCSR* system, which is based on subword unit search from word-based ASR, performs better on OOL queries than on INL queries. We consider this could be due to the larger OOV rate of the INL queries (18.2%) compared to that of the OOL queries (14.2%), which could affect the word-based ASR performance.

On RTVE data, the systems that only employ template matching approaches obtain, in general, better performance on OOL queries than on INL queries, which is due to the use of the query language (i.e., English). The only exceptions are the *F-Combined DTW* and *G-Super-BNF DTW* systems, for which the better MTWV performance

**Table 14** System results of the ALBAYZIN 2018 QbE STD evaluation on MAVIR test data for in-language (INL) and out-of-language (foreign) (OOL) queries

System ID	INL				OOL			
	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$
A-Hybrid DTW+LVCSR	0.2099	0.0997	0.00009	0.704	0.4918	0.4722	0.00008	0.430
B-Fusion DTW	0.2758	0.2714	0.00006	0.663	0.4836	0.4176	0.00016	0.359
C-PhonePost DTW	0.2322	0.2239	0.00003	0.740	0.4079	0.4079	0.00002	0.572
D-LVCSR	0.1454	0.1019	0.00008	0.778	0.2464	0.2268	0.00004	0.714
E-DTW	0.1458	0.1458	0.00001	0.849	0.2847	0.2847	0.00000	0.715
F-Combined DTW	0.1227	0.1165	0.00002	0.861	0.1643	0.1036	0.00016	0.679
G-Super-BNF DTW	0.1035	0.0324	0.00003	0.863	0.1554	−0.0089	0.00006	0.786
H-Multilingual-BNF DTW	0.1298	−0.0063	0.00005	0.817	0.2071	0.0500	0.00008	0.714
I-Monoph.-BNF DTW	0.1307	0.0848	0.00002	0.846	0.1429	0.0250	0.00000	0.857
J-Triph.-BNF DTW	0.1193	0.0658	0.00004	0.836	0.1429	0.0179	0.00000	0.857
K-Text STD	0.5647	0.5469	0.00011	0.327	0.1911	0.1071	0.00006	0.750

**Table 15** System results of the ALBAYZIN 2018 QbE STD evaluation on RTVE test data for in-language (INL) and out-of-language (foreign) (OOL) queries

System ID	INL				OOL			
	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$
A-Hybrid DTW+LVCSR	0.6502	0.5887	0.00008	0.268	0.5017	0.4821	0.00013	0.368
B-Fusion DTW	0.4126	0.3603	0.00004	0.549	0.5765	0.5043	0.00002	0.406
C-PhonePost DTW	0.3972	0.3175	0.00002	0.586	0.5319	0.4447	0.00000	0.464
D-LVCSR	0.5289	0.4560	0.00008	0.391	0.2155	0.1236	0.00007	0.719
E-DTW	0.0000	-0.0126	0.00000	1.000	0.0000	-0.0207	0.00000	1.000
F-Combined DTW	0.3193	0.3059	0.00002	0.658	0.3484	0.3058	0.00002	0.629
G-Super-BNF DTW	0.3026	0.3012	0.00003	0.665	0.3077	0.2946	0.00004	0.655
H-Multilingual-BNF DTW	0.2909	0.2841	0.00005	0.664	0.3104	0.2996	0.00013	0.555
I-Monoph.-BNF DTW	0.3179	0.3176	0.00005	0.629	0.3593	0.3519	0.00003	0.607
J-Triph.-BNF DTW	0.3123	0.2959	0.00003	0.654	0.3252	0.2968	0.00001	0.669
K-Text STD	0.7341	0.7189	0.00004	0.221	0.5924	0.4324	0.00008	0.326

on OOL queries indicates some threshold calibration issues, and the *E-DTW* system, for which the performance ( $\text{ATWV} < 0$ ) is meaningless. For the systems that employ ASR (i.e., *A-Hybrid DTW+LVCSR* and *D-LVCSR* systems), the performance is better on INL queries, since the ASR language matches that of the query.

On COREMAH data, systems obtain, in general, better MTWV performance on OOL queries than on INL queries, which is due to the use of the English language for system construction. Threshold calibration issues lead to higher ATWV for INL queries in some cases. For the *D-LVCSR* system, which is based on ASR, the MTWV performance is better on INL queries than on OOL

queries, which is consistent with the match between the language of the ASR system and the queries. However, threshold calibration issues produce a worse ATWV on INL queries.

As expected, the *K-Text STD* system, which is language-dependent and relies on the search in word lattices output by a Spanish ASR system, obtains better performance on INL queries than on OOL queries, since the query language matches the ASR target language. The only exception is the COREMAH data, for which a better MTWV performance on INL queries suggests threshold calibration issues in domains for which no development data are provided.

**Table 16** System results of the ALBAYZIN 2018 QbE STD evaluation on COREMAH test data for in-language (INL) and out-of-language (foreign) (OOL) queries

System ID	INL				OOL			
	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$
A-Hybrid DTW+LVCSR	0.1325	-0.0723	0.00006	0.806	0.2667	0.0840	0.00000	0.733
B-Fusion DTW	0.2008	-1.8806	0.00000	0.799	0.2667	-3.7685	0.00000	0.733
C-PhonePost DTW	0.2045	-1.8397	0.00000	0.795	0.2667	-4.8663	0.00000	0.733
D-LVCSR	0.0037	-0.1594	0.00000	0.996	0.0000	0.0000	0.00000	1.000
E-DTW	0.1512	0.1512	0.00002	0.828	0.0000	-0.1989	0.00000	1.000
F-Combined DTW	0.0532	-0.0014	0.00000	0.947	0.1667	0.1667	0.00000	0.833
G-Super-BNF DTW	0.0302	-0.4908	0.00000	0.970	0.1667	0.1667	0.00000	0.833
H-Multilingual-BNF DTW	0.0526	-0.9091	0.00000	0.947	0.1667	-0.0160	0.00000	0.833
I-Monoph.-BNF DTW	0.0512	-0.2715	0.00002	0.932	0.1667	-0.5640	0.00000	0.833
J-Triph.-BNF DTW	0.0585	-0.4515	0.00000	0.937	0.1667	0.1667	0.00000	0.833
K-Text STD	0.0988	-0.5918	0.00007	0.831	0.0000	-0.1828	0.00000	1.000

## 6.2 Performance analysis of QbE STD systems for single and multi-word queries

A similar analysis has been carried out for single and multi-word queries, and the results are presented in Tables 17, 18, and 19 for MAVIR, RTVE, and COREMAH databases, respectively. They show that system performance on multi-word queries is always better than on single-word queries for MAVIR and RTVE databases. We consider this is due to the fact that multi-word queries are longer than single-word queries, and hence produce less FAs so that the final performance gets improved.

On COREMAH data, for which no development data were provided, the performance drops dramatically. In addition, there is just one multi-word query, for which no detections are given by any system. The single-word query detection fails in threshold calibration in most of the cases, hence obtaining an ATWV  $< 0$ . The only system that obtains an ATWV  $> 0$  is the *E-DTW* system due to a perfect threshold calibration. This perfect calibration may be due to the fact that MAVIR development data were used for parameter tuning in the experiments using COREMAH data. MAVIR data present highly spontaneous and medium-quality speech data, which matches in some extent the speech of COREMAH data. On the other hand, RTVE data, which were used for parameter tuning in the rest of the systems (except for the *F-Combined DTW*, *G-Super-BNF DTW*, *H-Multilingual-BNF DTW*, *I-Monoph.-BNF DTW*, and *J-Triph.-BNF DTW* systems, which employed MAVIR data as well), present well-pronounced and high-quality speech, which do not match COREMAH data, and hence degrades the performance. For those systems, the low performance may be due to the type of tuning carried out.

## 6.3 Performance analysis of QbE STD systems for INV and OOV queries

An analysis of the QbE STD systems and the *K-Text STD* system for in-vocabulary and out-of-vocabulary queries has been carried out, and the results are presented in Tables 20, 21, and 22 for MAVIR, RTVE, and COREMAH databases, respectively. They show that, for MAVIR and RTVE databases, the performance on INV queries is better than on OOV queries. Although many of the QbE STD systems presented do not rely on ASR (the only exceptions are the *D-LVCSR* and the *A-Hybrid DTW+LVCSR* systems), system performance is, theoretically, better on INV queries than on OOV queries, due to the different properties INV and OOV queries convey. However, on COREMAH data, the MTWV obtained on OOV queries is in general, better than on INV queries. Since no development data were provided for this database, INV and OOV query detection must rely on parameter tuning that does not match the data domain, making INV query detection more difficult. The performance gaps between MTWV and ATWV metrics suggest some threshold calibration issues on COREMAH data, due to the lack of development data for this domain.

As expected, the *K-Text STD* system obtains better performance on INV queries for all the databases due to the match in the target language and the presence of the query terms in the vocabulary of the ASR system.

## 6.4 System fusion

After the evaluation, we have tried to combine 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.

**Table 17** System results of the ALBAYZIN 2018 QbE STD evaluation on MAVIR test data for single-word (Single) and multi-word (Multi) queries

System ID	Single				Multi			
	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$
A-Hybrid DTW+LVCSR	0.1905	0.0919	0.00017	0.637	0.7562	0.6646	0.00016	0.083
B-Fusion DTW	0.2586	0.2543	0.00007	0.673	0.7500	0.7271	0.00000	0.250
C-PhonePost DTW	0.2146	0.2052	0.00003	0.759	0.7500	0.7500	0.00000	0.250
D-LVCSR	0.1239	0.0800	0.00008	0.795	0.6437	0.6125	0.00002	0.333
E-DTW	0.1356	0.1356	0.00000	0.860	0.5000	0.4771	0.00000	0.500
F-Combined DTW	0.0945	0.0872	0.00004	0.863	0.7271	0.5896	0.00002	0.250
G-Super-BNF DTW	0.0788	0.0112	0.00008	0.839	0.7271	0.3375	0.00002	0.250
H-Multilingual-BNF DTW	0.0995	-0.0244	0.00003	0.872	0.6437	0.3604	0.00002	0.333
I-Monoph.-BNF DTW	0.0958	0.0545	0.00002	0.882	0.7271	0.5208	0.00002	0.250
J-Triph.-BNF DTW	0.0886	0.0475	0.00004	0.873	0.7042	0.3146	0.00005	0.250
K-Text STD	0.5316	0.5139	0.00011	0.361	0.7500	0.5833	0.00000	0.250

**Table 18** System results of the ALBAYZIN 2018 QbE STD evaluation on RTVE test data for single-word (Single) and multi-word (Multi) queries

System ID	Single				Multi			
	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$
A-Hybrid DTW+LVCSR	0.6122	0.5651	0.00008	0.310	0.9531	0.7374	0.00005	0.000
B-Fusion DTW	0.4369	0.3849	0.00004	0.523	0.4656	0.4093	0.00000	0.534
C-PhonePost DTW	0.4167	0.3374	0.00002	0.565	0.4562	0.4281	0.00001	0.534
D-LVCSR	0.4565	0.3854	0.00008	0.460	0.9906	0.8218	0.00001	0.000
E-DTW	0.0000	-0.0145	0.00000	1.000	0.0000	0.0000	0.00000	1.000
F-Combined DTW	0.3066	0.2904	0.00003	0.664	0.8860	0.8485	0.00002	0.095
G-Super-BNF DTW	0.2879	0.2849	0.00004	0.670	0.8672	0.8290	0.00004	0.095
H-Multilingual-BNF DTW	0.2765	0.2724	0.00005	0.677	0.8478	0.7915	0.00001	0.143
I-Monoph.-BNF DTW	0.3073	0.3070	0.00005	0.642	0.9055	0.9055	0.00005	0.048
J-Triph.-BNF DTW	0.2960	0.2808	0.00004	0.668	0.8583	0.8302	0.00001	0.132
K-Text STD	0.6917	0.6625	0.00008	0.229	0.8783	0.8783	0.00000	0.122

#### 6.4.1 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 query detection output by a certain system *A*, in case some other fused system *B* does not detect it (and hence, the corresponding score does not exist for it), the score generated for that detection is the minimum global score for system *B*.

#### 6.4.2 Calibration and fusion

Calibration and fusion are carried out with the Bosaris toolkit [123]. To do so, a linear model based on logistic regression trained from the scores of the detections of development queries is employed. MAVIR and RTVE data parameters are optimized independently from their

corresponding development sets and then applied to their corresponding test sets. For COREMAH data, the model trained for MAVIR data is employed.

Fusion employs the three primary systems corresponding to the three participants in the evaluation (i.e., *E-DTW*, *A-Hybrid DTW+LVCSR*, and *F-Combined DTW* systems).

#### 6.4.3 Fusion results

The results of the primary system fusion are presented in Table 23 for development data and Table 24 for test data. They show that system fusion enhances the performance of the best individual QbE STD system on MAVIR and COREMAH data, and the opposite stands for RTVE data. A paired *t* test shows that the best performance of the *Fusion* system is statistically significant ( $p <$

**Table 19** System results of the ALBAYZIN 2018 QbE STD evaluation on COREMAH test data for single-word (Single) and multi-word (Multi) queries

System ID	Single				Multi			
	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$
A-Hybrid DTW+LVCSR	0.1363	-0.0580	0.00006	0.803	0.0000	0.0000	0.00000	1.000
B-Fusion DTW	0.2038	-1.9448	0.00000	0.796	0.0000	0.0000	0.00000	1.000
C-PhonePost DTW	0.2075	-1.9184	0.00000	0.792	0.0000	0.0000	0.00000	1.000
D-LVCSR	0.0037	-0.1454	0.00000	0.996	0.0000	0.0000	0.00000	1.000
E-DTW	0.1445	0.1445	0.00003	0.827	0.0000	0.0000	0.00000	1.000
F-Combined DTW	0.0526	0.0023	0.00000	0.947	0.0000	0.0000	0.00000	1.000
G-Super-BNF DTW	0.0299	-0.4707	0.00000	0.970	0.0000	0.0000	0.00000	1.000
H-Multilingual-BNF DTW	0.0520	-0.8796	0.00000	0.948	0.0000	0.0000	0.00000	1.000
I-Monoph.-BNF DTW	0.0544	-0.2688	0.00002	0.929	0.0000	0.0000	0.00000	1.000
J-Triph.-BNF DTW	0.0578	-0.4231	0.00000	0.938	0.0000	0.0000	0.00000	1.000
K-Text STD	0.0977	-0.5893	0.00007	0.833	0.0000	0.0000	0.00000	1.000

**Table 20** System results of the ALBAYZIN 2018 QbE STD evaluation on MAVIR test data for in-vocabulary (INV) and out-of-vocabulary (OOV) queries

System ID	INV				OOV			
	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$
A-Hybrid DTW+LVCSR	0.2529	0.1586	0.00017	0.580	0.0709	-0.1441	0.00007	0.860
B-Fusion DTW	0.3057	0.3011	0.00006	0.636	0.2063	0.1239	0.00000	0.794
C-PhonePost DTW	0.2531	0.2439	0.00003	0.719	0.1811	0.1749	0.00003	0.784
D-LVCSR	0.1716	0.1335	0.00007	0.757	0.0680	-0.0728	0.00002	0.909
E-DTW	0.1673	0.1673	0.00000	0.830	0.0583	0.0583	0.00002	0.918
F-Combined DTW	0.1323	0.1261	0.00004	0.827	0.0735	0.0337	0.00006	0.869
G-Super-BNF DTW	0.1122	0.0367	0.00002	0.866	0.0749	-0.0257	0.00001	0.914
H-Multilingual-BNF DTW	0.1352	0.0074	0.00006	0.800	0.0688	-0.0807	0.00001	0.920
I-Monoph.-BNF DTW	0.1417	0.0941	0.00002	0.835	0.0459	-0.0228	0.00001	0.943
J-Triph.-BNF DTW	0.1308	0.0694	0.00004	0.825	0.0451	0.0094	0.00001	0.943
K-Text STD	0.5639	0.5542	0.00010	0.338	0.3582	0.2327	0.00022	0.424

0.01) compared to the best QbE STD system on MAVIR test data (*A-Hybrid DTW+LVCSR*), and weakly significant ( $p < 0.08$ ) compared to the best QbE STD system on MAVIR development data (*A-Hybrid DTW+LVCSR*). This highlights the power of fused systems in QbE STD in challenging domains that include medium-quality and highly spontaneous speech data. The drop in performance of the *Fusion* system compared to the best QbE STD system on RTVE test data (*A-Hybrid DTW+LVCSR*) is not statistically significant for a paired  $t$  test.

The *K-Text STD* system performs better than the *Fusion* system for MAVIR and RTVE data. This improvement in performance is statistically significant for a paired  $t$  test ( $p < 0.01$ ) for both development and test sets of

MAVIR data and for the test set of RTVE data. However, on COREMAH test data, the *Fusion* system outperforms the *K-Text STD* system. This improvement in performance is statistically significant for a paired  $t$  test ( $p < 0.02$ ), which indicates that fusing QbE STD systems that are based on different strategies can outperform text-based STD technology on unseen data domains.

DET curves of the fusion systems along with the rest of the primary systems and the *K-Text STD* system are presented in Figs. 11, 12, 13, 14, and 15 for development/test MAVIR, RTVE, and COREMAH data. Comparing the QbE STD systems on MAVIR development and test data, it can be seen that (1) the *Fusion* system performs the best, except for very low FA rates, for which the *E-DTW*

**Table 21** System results of the ALBAYZIN 2018 QbE STD evaluation on RTVE test data for in-vocabulary (INV) and out-of-vocabulary (OOV) queries

System ID	INV				OOV			
	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$
A-Hybrid DTW+LVCSR	0.6948	0.6449	0.00007	0.233	0.3401	0.2590	0.00005	0.608
B-Fusion DTW	0.4525	0.3969	0.00004	0.507	0.3879	0.3388	0.00001	0.603
C-PhonePost DTW	0.4277	0.3540	0.00002	0.555	0.3837	0.2814	0.00002	0.595
D-LVCSR	0.5527	0.4920	0.00008	0.369	0.1674	0.0061	0.00003	0.800
E-DTW	0.0000	-0.0175	0.00000	1.000	0.0000	0.0000	0.00000	1.000
F-Combined DTW	0.3303	0.3091	0.00004	0.633	0.3072	0.2927	0.00010	0.591
G-Super-BNF DTW	0.3057	0.3033	0.00003	0.661	0.2932	0.2865	0.00003	0.672
H-Multilingual-BNF DTW	0.2970	0.2918	0.00004	0.661	0.2812	0.2663	0.00003	0.689
I-Monoph.-BNF DTW	0.3427	0.3424	0.00005	0.606	0.2737	0.2461	0.00009	0.637
J-Triph.-BNF DTW	0.3092	0.2956	0.00003	0.657	0.3218	0.2979	0.00004	0.638
K-Text STD	0.7896	0.7733	0.00003	0.179	0.4391	0.2344	0.00011	0.451

**Table 22** System results of the ALBAYZIN 2018 QbE STD evaluation on COREMAH test data for in-vocabulary (INV) and out-of-vocabulary (OOV) queries

System ID	INV				OOV			
	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$
A-Hybrid DTW+LVCSR	0.1265	-0.0821	0.00006	0.810	0.2215	0.0383	0.00004	0.742
B-Fusion DTW	0.1953	-1.8954	0.00000	0.805	0.2580	-2.1382	0.00000	0.742
C-PhonePost DTW	0.1994	-1.9037	0.00000	0.801	0.2580	-1.9260	0.00000	0.742
D-LVCSR	0.0071	-0.1616	0.00002	0.975	0.0000	-0.1096	0.00000	1.000
E-DTW	0.1458	0.1458	0.00002	0.832	0.1256	0.1256	0.00007	0.801
F-Combined DTW	0.0461	-0.0139	0.00000	0.954	0.1333	0.1333	0.00000	0.867
G-Super-BNF DTW	0.0301	-0.4915	0.00000	0.970	0.0885	-0.3533	0.00004	0.875
H-Multilingual-BNF DTW	0.0454	-0.9151	0.00000	0.955	0.1000	-0.6821	0.00000	0.900
I-Monoph.-BNF DTW	0.0440	-0.3012	0.00002	0.938	0.1333	-0.0897	0.00000	0.867
J-Triph.-BNF DTW	0.0519	-0.4771	0.00000	0.944	0.1833	-0.1205	0.00000	0.817
K-Text STD	0.1085	-0.3436	0.00008	0.814	0.0000	-2.5204	0.00000	1.000

system performs the best, and (2) the *K-Text STD* system performs better than any QbE STD system for all the operating points. On RTVE development data, the *Fusion* system performs the best, except for very low miss rates, for which the *A-Hybrid DTW+LVCSR* system performs the best. Comparing the QbE STD systems on RTVE test data, it can be seen that (1) the *Fusion* system performs the best, except for very low FA rates and low miss rates, for which the *A-Hybrid DTW+LVCSR* system performs the best, and (2) the *K-Text STD* system performs better than any QbE STD system for all the operating points. Comparing the QbE STD systems on COREMAH test data, it can be seen that (1) the *Fusion* system performs the best for all the operating points, except for very low FA rates, for which the *F-Combined DTW* system performs the best, and for very low miss rates, for which the *A-Hybrid DTW+LVCSR* system obtains the best performance, and (2) the *K-Text STD* system outperforms any QbE STD system in low miss rates.

These results highlight the power of fusing systems in QbE STD since the *Fusion* system obtains, in general, the best performance across the different datasets, and in

some scenarios, QbE STD outperforms text-based STD using textual queries.

### 6.5 Comparison to the ALBAYZIN 2016 QbE STD evaluation

The evaluations carried out in 2016 and 2018 share the MAVIR data (queries and utterances). Therefore, a comparison between the best system submitted to both evaluations can be carried out. On MAVIR test data, the best result obtained in the 2018 evaluation is  $\text{ATWV} = 0.2810$ , which is higher than that obtained in the previous ALBAYZIN 2016 QbE STD evaluation ( $\text{ATWV} = 0.2646$ ). The best performance in 2016 corresponded to a combined system that integrated DTW search on different feature sets. However, in the 2018 evaluation, the detections obtained from the different feature sets are added the detections from a text STD approach (hence resulting in a hybrid QbE STD system). This hybrid system, which integrates two standard approaches for QbE STD, is clearly giving a better performance than systems that only integrate template matching approaches.

**Table 23** Fusion system results of the ALBAYZIN 2018 QbE STD evaluation on development data

System ID	MAVIR				RTVE			
	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$
A-Hybrid DTW+LVCSR	0.2896	0.2470	0.00010	0.606	0.7273	0.6964	0.00002	0.254
E-DTW	0.1823	0.1774	0.00002	0.801	0.0365	0.0365	0.00000	0.963
F-Combined DTW	0.1512	0.0938	0.00005	0.803	0.6003	0.5874	0.00006	0.339
Fusion	0.3354	0.3124	0.00012	0.540	0.7198	0.6489	0.00003	0.253
K-Text STD	0.6544	0.6042	0.00012	0.229	0.7591	0.6918	0.00006	0.184

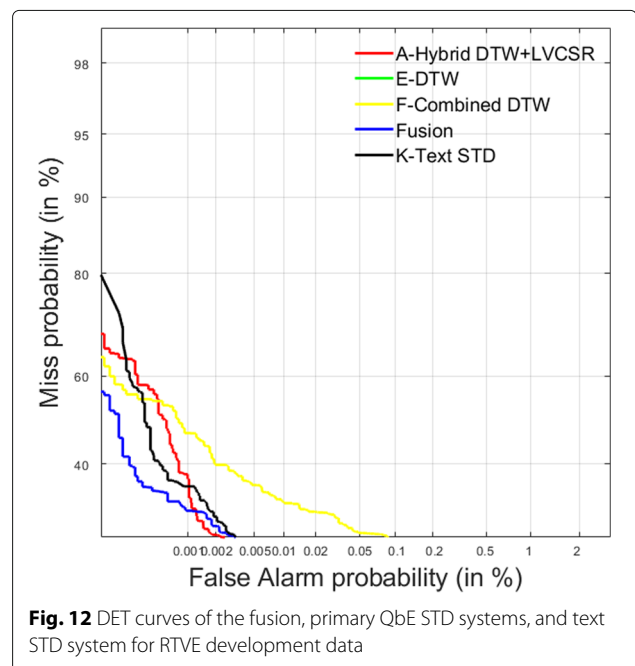
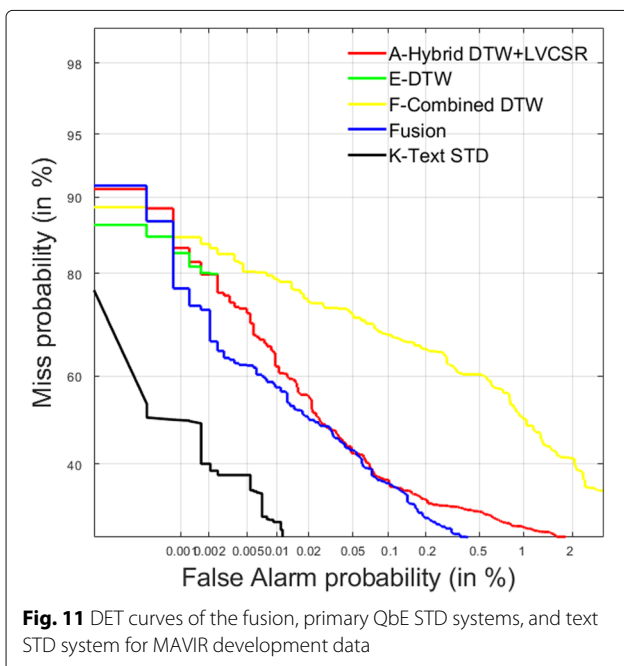
**Table 24** Fusion system results of the ALBAYZIN 2018 QbE STD evaluation on test data

System ID	MAVIR				RTVE			
	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$	MTWV	ATWV	$p(\text{FA})$	$p(\text{Miss})$
A-Hybrid DTW+LVCSR	0.2226	0.1243	0.00017	0.606	0.6201	0.5699	0.00008	0.301
E-DTW	0.1550	0.1550	0.00001	0.840	0.0000	− 0.0141	0.00000	1.000
F-Combined DTW	0.1227	0.1157	0.00002	0.860	0.3224	0.3059	0.00003	0.648
Fusion	0.2769	0.2700	0.00007	0.649	0.6216	0.5564	0.00004	0.340
K-Text STD	0.5345	0.5178	0.00010	0.364	0.6969	0.6685	0.00008	0.226
COREMAH								
A-Hybrid DTW+LVCSR	0.1354	− 0.0689	0.00006	0.804				
E-DTW	0.1436	0.1436	0.00003	0.828				
F-Combined DTW	0.0521	0.0023	0.00000	0.948				
Fusion	0.1779	0.1730	0.00001	0.810				
K-Text STD	0.0966	− 0.5828	0.00007	0.835				

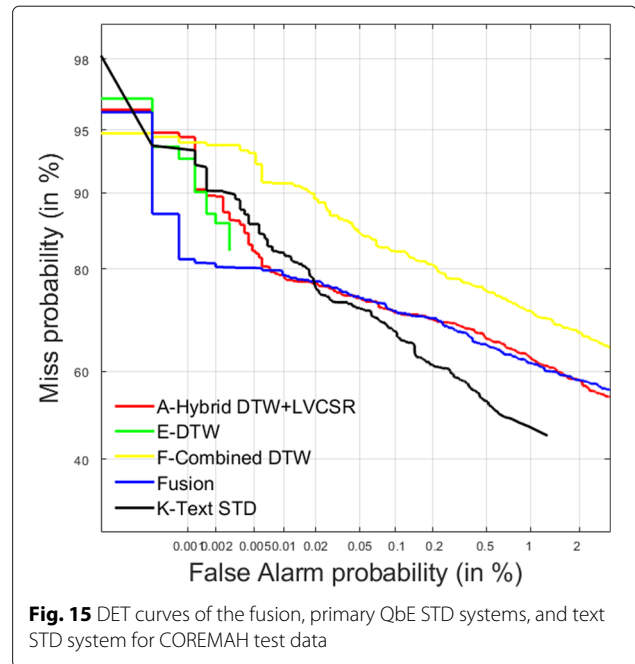
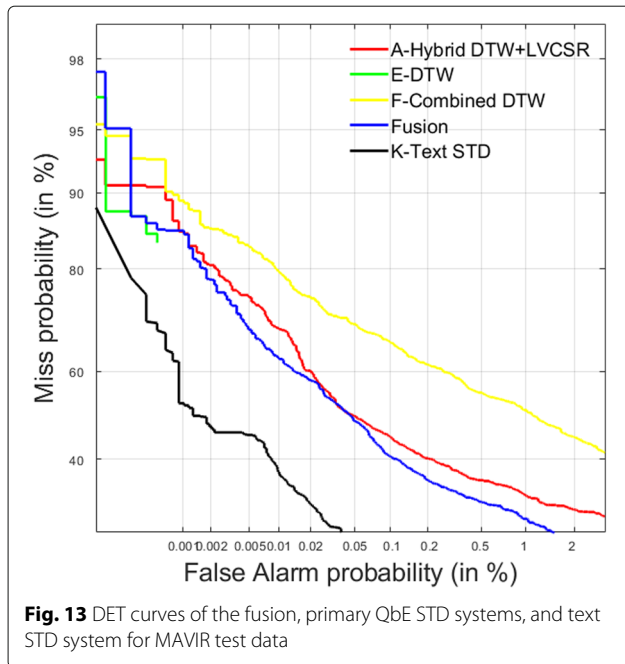
### 6.6 Towards a language-independent STD system

Due to the intrinsic language independence of various QbE STD systems submitted to this evaluation (see Table 6), the feasibility of language-independent STD systems can be examined. From the overall evaluation results (see Table 8), it can be seen that language-independent STD systems are still far from obtaining better or even similar performance to that obtained with language-dependent STD systems. The performance obtained with the best language-independent system (i.e., *B-Fusion DTW*) is  $\text{ATWV} = 0.3082$ , and the performance

obtained with the *K-Text STD* system is  $\text{ATWV} = 0.4427$ , which suggests that language-independent STD still represents a challenge. This is clearer for domains in which training/development data are given in advance for system training and tuning (see Tables 11 and 12). When the data domain changes (as COREMAH data in this evaluation), the performance of language-dependent STD systems drops dramatically so that language-independent STD systems may obtain similar or even better performance compared to language-dependent STD systems (see Table 13). Therefore, it can be claimed that language-







independent STD systems are feasible for out-of-domain data.

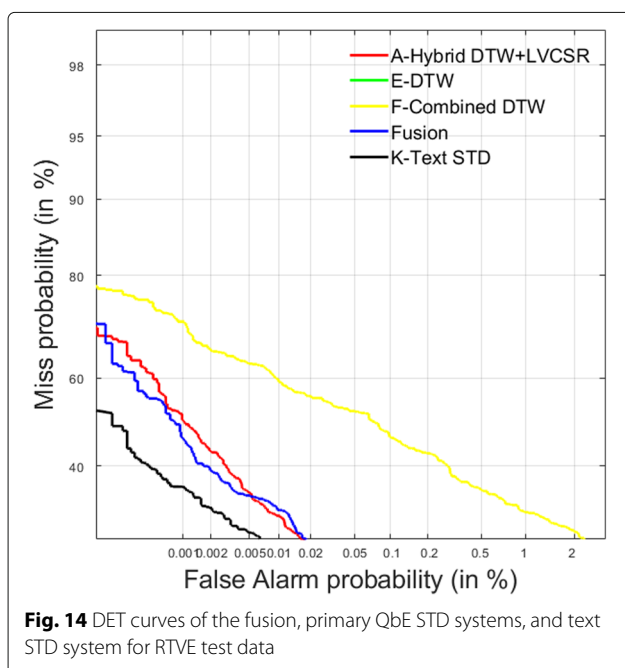
## 7 Conclusions

This paper has presented a multi-domain international QbE STD evaluation for SoS in Spanish. The amount of systems submitted to the evaluation has made it possible to compare the progress of QbE STD technology

under a common framework. Three different teams participated in the evaluation and ten different systems were submitted. Additionally, a text-based STD system has also been presented to compare STD and QbE STD technologies. Systems belong to three well-known QbE STD categories: text-based STD, template matching, and hybrid. Among those systems, *A-Hybrid DTW+LVCSR* and *D-LVCSR* systems, which include a probabilistic retrieval model for information retrieval and a query likelihood retrieval model, and *F-Combined DTW*, *G-Super-BNF DTW*, *H-Multilingual-BNF DTW*, *I-Monoph.-BNF DTW*, and *J-Triph.-BNF DTW*, which employ stacked bottleneck features for signal representation, can be considered novel from a QbE STD perspective.

Results have shown a high variability with regard to domain change. On the one hand, systems have obtained the best performance on RTVE data, for which a large amount of training data are available for system construction and present high-quality and well-pronounced speech. For these data, hybrid systems are typically the best choice due to those afore-mentioned characteristics. On the other hand, systems have obtained the worst performance on COREMAH data, for which only test data were provided. This indicates that domain change is quite challenging in QbE STD. On MAVIR data, which are also quite challenging due to the presence of spontaneous speech, system performance was between those for RTVE and COREMAH data.

We have also shown that template matching systems for which the language of the foreign queries is employed in development (e.g., for feature extraction) obtained



better performance on OOL query detection than on INL query detection. Systems have obtained better performance on multi-word query detection than on single-word query detection because lower FA rates are generally obtained on longer queries. Systems have obtained better performance on INV queries than on OOV queries for domains for which development data are provided, since OOV queries convey, in general, more diverse properties. However, for out-of-domain data, system performance on OOV queries may be better than on INV queries since the change in the data domain is more critical, especially for the systems based on template matching.

Given the best overall result obtained in the evaluation (ATWV = 0.3260), which comes from the average of the three domains, there is still an ample room for improvement. Specifically, it has been observed that QbE STD systems degrade to a great extent in unseen data domains, for which language-independent STD systems (ATWV = 0.1436) outperformed language-dependent STD systems (ATWV = -0.5828). This encourages us to maintain the QbE STD evaluation in the next years, focusing on multi-domain QbE STD.

## Endnotes

<sup>1</sup> <http://www.rthabla.es/>

<sup>2</sup> <http://www.isca-speech.org/iscaweb/index.php/signs?layout=edit&id=132>

<sup>3</sup> <http://www.mavir.net>

<sup>4</sup> <http://cartago.lllf.uam.es/mavir/index.pl?m=videos>

<sup>5</sup> <http://sox.sourceforge.net/>

<sup>6</sup> <http://www.lllf.uam.es/coremah/>

<sup>7</sup> <https://ffmpeg.org/>

<sup>8</sup> <http://lucene.apache.org>

<sup>9</sup> <http://www.tc-star.org>

<sup>10</sup> <http://cartago.lllf.uam.es/mavir/index.pl?m=descargas>

## Abbreviations

AAC: Advanced audio coding; ASR: Automatic speech recognition; ATWV: Actual term-weighted value; BUT: Brno University of Technology; DET: Detection error tradeoff; DNN: Deep neural network; DTW: Dynamic time warping; FA: False alarm; GMM: Gaussian mixture model; HMM: Hidden Markov model; HNR: Harmonics-to-noise ratio; IARPA: Intelligence advanced research projects activity; INL: In-language; INV: In-vocabulary; KWS: Keyword spotting; LM: Language model; LVCSR: Large vocabulary continuous speech recognition; MED: Minimum edit distance; MFCC: Mel-frequency cepstral coefficient; MOS: Mean opinion score; MPEG: Moving picture experts group; MTWV: Maximum term-weighted value; NIST: National institute of standards and technology; NS-DTW: Non-segmental dynamic time warping; OOL: Out-of-language; OOV: Out-of-vocabulary; PCM: Pulse code modulation; QbE STD: Query-by-Example Spoken Term Detection; QUESST: Query-by-Example Search on Speech Task; RTVE: Radio Televisión Española; S-DTW: Subsequence DTW; sBNF: Stacked bottleneck feature; SDR: Spoken document retrieval; SIG-IL: Special interest group on iberian languages; SoS: Search on speech; STD: Spoken term detection; SWS: Spoken web search; TV: Television; TWV: Term-weighted value; VAD: Voice activity detection; WFST: Weighted finite state transducer

## Authors' contributions

JT and DTT designed and prepared the QbE STD evaluation, built the *E-DTW* system, and carried out the post-evaluation analysis. PL-O and LD-F built the *A-Hybrid DTW+LVCSR*, *B-Fusion DTW*, *C-PhonePost DTW*, *D-LVCSR*, and *K-Text STD* systems. MP and LJR-F built the *F-Combined DTW*, *G-Super-BNF DTW*, *H-Multilingual-BNF DTW*, *I-Monoph.-BNF DTW*, and *J-Triph.-BNF DTW* systems and carried out the primary system fusion. AM-S provided the MAVIR and COREMAH databases, collaborated with labeling the new data for the evaluation, and provided linguistic support. All the authors contributed in the final discussion of the results. The main contributions of this paper are as follows: (1) Systems submitted to the fourth Query-by-Example Spoken Term Detection evaluation for Spanish language are presented. (2) A 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, Xunta de Galicia - "Consellería de Cultura, Educación e Ordenación Universitaria", the European Regional Development Fund through the 2016-2019 accreditations ED431G/01 ("Centro singular de investigación de Galicia") and ED431G/04 ("Agrupación estratégica consolidada"), the UPV/EHU under grant GIU16/68, 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), the project "DSForSec: Deep Speech for Forensics and Security" funded by the Ministry of Science, Innovation and Universities and FEDER (RTI2018-098091-B-I00), and the Cátedra UNIZAR-RTVE.

## Availability of data and materials

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

## Author details

<sup>1</sup>Escuela Politécnica Superior, Fundación Universitaria San Pablo CEU, Campus de Montepíncipe, Madrid, Spain. <sup>2</sup>AUDIAS, Universidad Autónoma de Madrid, Av. Francisco Tomás y Valiente, 11. Escuela Politécnica Superior, Madrid, Spain. <sup>3</sup>Universidad da Coruña, IRLab, CITIC, Campus de Elviña s/n, A Coruña, Spain. <sup>4</sup>Multimedia Technologies Group (GTM), AtlantTIC Research Center, E. E. Telecomunicación, Campus Universitario de Vigo, s/n, Vigo, Spain. <sup>5</sup>Software Technology Working Group (GTTS), Universidad del País Vasco, Barrio Sarriena s/n, Leioa, Spain. <sup>6</sup>Laboratorio de Lingüística Informática, Universidad Autónoma de Madrid, Carretera de Colmenar, km. 16. Facultad de Filosofía y Letras, Madrid, Spain.

Received: 12 April 2019 Accepted: 26 June 2019

Published online: 19 July 2019

## References

1. K. Ng, V. W. Zue, Subword-based approaches for spoken document retrieval. *Speech Commun.* **32**(3), 157–186 (2000)
2. 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), 2602–2612 (2012)
3. T.-H. Lo, Y.-W. Chen, K.-Y. Chen, H.-M. Wang, B. Chen, in *Proc. of ASRU*. Neural relevance-aware query modeling for spoken document retrieval (IEEE, USA, 2017), pp. 466–473
4. W. F. L. Heeren, F. M. G. de Jong, L. B. van der Werff, M. A. H. Huijbregts, R. J. F. Ordelman, in *Proc. of LREC*. Evaluation of spoken document retrieval for historic speech collections (ELRA, Belgium, 2008), pp. 2037–2041
5. 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)
6. Y.-W. Chen, K.-Y. Chen, H.-M. Wang, B. Chen, in *Proc. of Interspeech*. Exploring the use of significant words language modeling for spoken document retrieval (ISCA, France, 2017), pp. 2889–2893

7. P. Gao, J. Liang, P. Ding, B. Xu, in *Proc. of ICASSP*. A novel phone-state matrix based vocabulary-independent keyword spotting method for spontaneous speech (IEEE, USA, 2007), pp. 425–428
8. B. Zhang, R. Schwartz, S. Tsakalidis, L. Nguyen, S. Matsoukas, in *Proc. of Interspeech*. White listing and score normalization for keyword spotting of noisy speech (ISCA, France, 2012), pp. 1832–1835
9. 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 *Proc. of Interspeech*. Strategies for high accuracy keyword detection in noisy channels (ISCA, France, 2013), pp. 15–19
10. T. Ng, R. Hsiao, L. Zhang, D. Karakos, S. H. Mallidi, M. Karafiat, K. Vesely, I. Szoke, B. Zhang, L. Nguyen, R. Schwartz, in *Proc. of Interspeech*. Progress in the BBN keyword search system for the DARPA RATS program (ISCA, France, 2014), pp. 959–963
11. V. Mitra, J. van Hout, H. Franco, D. Vergyri, Y. Lei, M. Graciarena, Y.-C. Tam, J. Zheng, in *Proc. of ICASSP*. Feature fusion for high-accuracy keyword spotting (IEEE, USA, 2014), pp. 7143–7147
12. S. Panchapagesan, M. Sun, A. Khare, S. Matsoukas, A. Mandal, B. Hoffmeister, S. Vitaladevuni, in *Proc. of Interspeech*. Multi-task learning and weighted cross-entropy for DNN-based keyword spotting (ISCA, France, 2016), pp. 760–764
13. J. Mamou, B. Ramabhadran, O. Siohan, in *Proc. of ACM SIGIR*. Vocabulary independent spoken term detection (ACM, USA, 2007), pp. 615–622
14. D. Schneider, T. Mertens, M. Larson, J. Kohler, in *Proc. of Interspeech*. Contextual verification for open vocabulary spoken term detection (ISCA, France, 2010), pp. 697–700
15. C. Parada, A. Sethy, M. Dredze, J. Jelinek, in *Proc. of Interspeech*. A spoken term detection framework for recovering out-of-vocabulary words using the web (ISCA, France, 2010), pp. 1269–1272
16. I. Szöke, M. Fapšo, L. Burget, J. Černocký, in *Proc. of Speech Search Workshop at SIGIR*. Hybrid word-subword decoding for spoken term detection (ACM, USA, 2008), pp. 42–48
17. Y. Wang, F. Metzger, in *Proc. of Interspeech*. An in-depth comparison of keyword specific thresholding and sum-to-one score normalization (ISCA, France, 2014), pp. 2474–2478
18. L. Mangu, G. Saon, M. Picheny, B. Kingsbury, in *Proc. of ICASSP*. Order-free spoken term detection (IEEE, USA, 2015), pp. 5331–5335
19. A. Buzo, H. Cucu, C. Burileanu, in *Proc. of MediaEval*. Speed@MediaEval 2014: Spoken term detection with robust multilingual phone recognition (CEUR, Germany, 2014), pp. 721–722
20. R. Konno, K. Ouchi, M. Obara, Y. Shimizu, T. Chiba, T. Hirota, Y. Itoh, in *Proc. 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
21. R. Jarina, M. Kuba, R. Gubka, M. Chmulik, M. Paralic, in *Proc. of MediaEval*. UNIZA system for the spoken web search task at MediaEval 2013 (CEUR, Germany, 2013), pp. 791–792
22. X. Anguera, M. Ferrarons, in *Proc. of ICME*. Memory efficient subsequence DTW for query-by-example spoken term detection (IEEE, USA, 2013), pp. 1–6
23. H. Lin, A. Stupakov, J. Bilmes, in *Proc. of Interspeech*. Spoken keyword spotting via multi-lattice alignment (ISCA, France, 2008), pp. 2191–2194
24. C. Chan, L. Lee, in *Proc. of Interspeech*. Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping (ISCA, France, 2010), pp. 693–696
25. S. Settle, K. Levin, H. Kamper, K. Livescu, in *Proc. of Interspeech*. Query-by-example search with discriminative neural acoustic word embeddings (ISCA, France, 2017), pp. 2874–2878
26. R. Shankar, C. M. Vikram, S. R. M. Prasanna, in *Proc. of Interspeech*. Spoken keyword detection using joint DTW-CNN (ISCA, France, 2018), pp. 117–121
27. A. Ali, M. A. Clements, in *Proc. of MediaEval*. Spoken web search using and ergodic hidden Markov model of speech (CEUR, Germany, 2013), pp. 861–862
28. A. Caranica, A. Buzo, H. Cucu, C. Burileanu, in *Proc. of MediaEval*. Speed@MediaEval 2015: Multilingual phone recognition approach to Query By Example STD (CEUR, Germany, 2015), pp. 781–783
29. S. Kesiraju, G. Mantena, K. Prahallad, in *Proc. of MediaEval*. IIIT-H system for MediaEval 2014 QUESST (CEUR, Germany, 2014), pp. 761–762
30. M. Ma, A. Rosenberg, in *Proc. of MediaEval*. CUNY systems for the Query-by-Example search on speech task at MediaEval 2015 (CEUR, Germany, 2015), pp. 831–833
31. J. Takahashi, T. Hashimoto, R. Konno, S. Sugawara, K. Ouchi, S. Oshima, T. Akyu, Y. Itoh, in *Proc. of NTCIR-11*. An IWAPU STD system for OOV query terms and spoken queries (Japan Society for Promotion of Science, Japan, 2014), pp. 384–389
32. M. Makino, A. Kai, in *Proc. of NTCIR-11*. Combining subword and state-level dissimilarity measures for improved spoken term detection in NTCIR-11 SpokenQuery&Doc task (Japan Society for Promotion of Science, Japan, 2014), pp. 413–418
33. N. Sakamoto, K. Yamamoto, S. Nakagawa, in *Proc. of ASRU*. Combination of syllable based N-gram search and word search for spoken term detection through spoken queries and IV/OOV classification (IEEE, USA, 2015), pp. 200–206
34. J. Hou, V. T. Pham, C.-C. Leung, L. Wang, H. Xu, H. Lv, L. Xie, Z. Fu, C. Ni, X. Xiao, H. Chen, S. Zhang, S. Sun, Y. Yuan, P. Li, T. L. Nwe, S. Sivasdas, B. Ma, E. S. Chng, H. Li, in *Proc. of MediaEval*. The NNI Query-by-Example system for MediaEval 2015 (IEEE, USA, 2015), pp. 141–143
35. J. Vavrek, P. Vizlay, M. Lojka, M. Pleva, J. Juhar, M. Rusko, in *Proc. of MediaEval*. TUKE at MediaEval 2015 QUESST (CEUR, Germany, 2015), pp. 451–453
36. H. Wang, T. Lee, C.-C. Leung, B. Ma, H. Li, Acoustic segment modeling with spectral clustering methods. *IEEE/ACM Trans. Audio Speech Lang. Process.* **23**(2), 264–277 (2015)
37. C.-T. Chung, L.-S. Lee, Unsupervised discovery of structured acoustic tokens with applications to spoken term detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **26**(2), 394–405 (2018)
38. C.-T. Chung, C.-Y. Tsai, C.-H. Liu, L.-S. Lee, Unsupervised iterative deep learning of speech features and acoustic tokens with applications to spoken term detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **25**(10), 1914–1928 (2017)
39. 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)
40. G. Mantena, S. Achanta, K. Prahallad, Query-by-example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(5), 946–955 (2014)
41. H. Tulsiani, P. Rao, in *Proc. of MediaEval*. The IIT-B Query-by-Example system for MediaEval 2015 (CEUR, Germany, 2015), pp. 341–343
42. M. Bouallegue, G. Senay, M. Morchid, D. Matrouf, G. Linares, R. Dufour, in *Proc. of MediaEval*. LIA@MediaEval 2013 spoken web search task: an I-vector based approach (CEUR, Germany, 2013), pp. 771–772
43. L. J. Rodriguez-Fuentes, A. Varona, M. Penagarikano, G. Bordel, M. Diez, in *Proc. of MediaEval*. GTTS systems for the SWS task at MediaEval 2013 (CEUR, Germany, 2013), pp. 831–832
44. H. Wang, T. Lee, C.-C. Leung, B. Ma, H. Li, in *Proc. of ICASSP*. Using parallel tokenizers with DTW matrix combination for low-resource spoken term detection (IEEE, USA, 2013), pp. 8545–8549
45. H. Wang, T. Lee, in *Proc. of MediaEval*. The CUHK spoken web search system for MediaEval 2013 (CEUR, Germany, 2013), pp. 681–682
46. J. Proenca, A. Veiga, F. Perdigão, in *Proc. of MediaEval*. The SPL-IT query by example search on speech system for MediaEval 2014 (CEUR, Germany, 2014), pp. 741–742
47. J. Proenca, A. Veiga, F. Perdigão, in *Proc. of EUSIPCO*. Query by example search with segmented dynamic time warping for non-exact spoken queries (Springer, Germany, 2015), pp. 1691–1695
48. J. Proenca, L. Castela, F. Perdigão, in *Proc. of MediaEval*. The SPL-IT-UC Query by Example search on speech system for MediaEval 2015 (CEUR, Germany, 2015), pp. 471–473
49. J. Proenca, F. Perdigão, in *Proc. of Interspeech*. Segmented dynamic time warping for spoken Query-by-Example search (ISCA, France, 2016), pp. 750–754
50. P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, in *Proc. of MediaEval*. GTM-UVigo systems for the Query-by-Example search on speech task at MediaEval 2015 (CEUR, Germany, 2015), pp. 521–523
51. P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, in *Proc. of ASRU*. Phonetic unit selection for cross-lingual Query-by-Example spoken term detection (IEEE, USA, 2015), pp. 223–229
52. A. Saxena, B. Yegnanarayana, in *Proc. of Interspeech*. Distinctive feature based representation of speech for Query-by-Example spoken term detection (ISCA, France, 2015), pp. 3680–3684
53. P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, in *Proc. of Interspeech*. Compensating gender variability in query-by-example

- search on speech using voice conversion (ISCA, France, 2017), pp. 2909–2913
54. A. Asaei, D. Ram, H. Bourlard, in *Proc. of Interspeech*. Phonological posterior hashing for query by example spoken term detection (ISCA, France, 2018), pp. 2067–2071
  55. M. Skacel, I. Szöke, in *Proc. of MediaEval*. BUT QUESST 2015 system description (CEUR, Germany, 2015), pp. 721–723
  56. H. Chen, C.-C. Leung, L. Xie, B. Ma, H. Li, in *Proc. of Interspeech*. Unsupervised bottleneck features for low-resource Query-by-Example spoken term detection (ISCA, France, 2016), pp. 923–927
  57. Y. Yuan, C.-C. Leung, L. Xie, H. Chen, B. Ma, H. Li, in *Proc. of ICASSP*. Pairwise learning using multi-lingual bottleneck features for low-resource Query-by-Example spoken term detection (IEEE, USA, 2017), pp. 5645–5649
  58. J. van Hout, V. Mitra, H. Franco, C. Bartels, D. Vergyri, in *Proc. of ASRU*. Tackling unseen acoustic conditions in query-by-example search using time and frequency convolution for multilingual deep bottleneck features (IEEE, USA, 2017), pp. 48–54
  59. E. Yilmaz, J. van Hout, H. Franco, in *Proc. of ASRU*. Noise-robust exemplar matching for rescoring query-by-example search (IEEE, USA, 2017), pp. 1–7
  60. Y. Yuan, C.-C. Leung, L. Xie, H. Chen, B. Ma, H. Li, in *Proc. of ICASSP*. Pairwise learning using multi-lingual bottleneck features for low-resource query-by-example spoken term detection (IEEE, USA, 2017), pp. 5645–5649
  61. A. H. H. N. Torbati, J. Picone, in *Proc. of Interspeech*. A nonparametric bayesian approach for spoken term detection by example query (ISCA, France, 2016), pp. 928–932
  62. A. Popli, A. Kumar, in *Proc. of MMSP*. Query-by-example spoken term detection using low dimensional posteriorgrams motivated by articulatory classes (IEEE, USA, 2015), pp. 1–6
  63. P. Yang, C.-C. Leung, L. Xie, B. Ma, H. Li, in *Proc. of Interspeech*. Intrinsic spectral analysis based on temporal context features for query-by-example spoken term detection (ISCA, France, 2014), pp. 1722–1726
  64. B. George, A. Saxena, G. Mantena, K. Prahallad, B. Yegnanarayana, in *Proc. of Interspeech*. Unsupervised query-by-example spoken term detection using bag of acoustic words and non-segmental dynamic time warping (ISCA, France, 2014), pp. 1742–1746
  65. D. Ram, A. Asaei, H. Bourlard, Sparse subspace modeling for query by example spoken term detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **26**(6), 1126–1139 (2018)
  66. P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, Finding relevant features for zero-resource query-by-example search on speech. *Speech Commun.* **84**, 24–35 (2016)
  67. T. J. Hazen, W. Shen, C. M. White, in *Proc. of ASRU*. Query-by-example spoken term detection using phonetic posteriorgram templates (IEEE, USA, 2009), pp. 421–426
  68. A. Abad, R. F. Astudillo, I. Trancoso, in *Proc. of MediaEval*. The L2F spoken web search system for MediaEval 2013 (CEUR, Germany, 2013), pp. 851–852
  69. I. Szöke, M. Skácel, L. Burget, in *Proc. of MediaEval*. BUT QUESST 2014 system description (CEUR, Germany, 2014), pp. 621–622
  70. I. Szöke, L. Burget, F. Grézil, J. H. Černocký, L. Ondel, in *Proc. of ICASSP*. Calibration and fusion of query-by-example systems - BUT SWS 2013 (IEEE, USA, 2014), pp. 7849–7853
  71. A. Abad, L. J. Rodríguez-Fuentes, M. Penagarikano, A. Varona, G. Bordel, in *Proc. of Interspeech*. On the calibration and fusion of heterogeneous spoken term detection systems (ISCA, France, 2013), pp. 20–24
  72. P. Yang, H. Xu, X. Xiao, L. Xie, C.-C. Leung, H. Chen, J. Yu, H. Lv, L. Wang, S. J. Leow, B. Ma, E. S. Chng, H. Li, in *Proc. of MediaEval*. The NNI query-by-example system for MediaEval 2014 (CEUR, Germany, 2014), pp. 691–692
  73. C.-C. Leung, L. Wang, H. Xu, J. Hou, V. T. Pham, H. Lv, L. Xie, X. Xiao, C. Ni, B. Ma, E. S. Chng, H. Li, in *Proc. of Interspeech*. Toward high-performance language-independent Query-by-Example spoken term detection for MediaEval 2015: Post-Evaluation analysis (ISCA, France, 2016), pp. 3703–3707
  74. H. Xu, J. Hou, X. Xiao, V. T. Pham, C.-C. Leung, L. Wang, V. H. Do, H. Lv, L. Xie, B. Ma, E. S. Chng, H. Li, in *Proc. of ICASSP*. Approximate search of audio queries by using DTW with phone time boundary and data augmentation (IEEE, USA, 2016), pp. 6030–6034
  75. S. Oishi, T. Matsuba, M. Makino, A. Kai, in *Proc. of NTCIR-12*. Combining state-level and DNN-based acoustic matches for efficient spoken term detection in NTCIR-12 SpokenQuery&Doc-2 task (Japan Society for Promotion of Science, Japan, 2016), pp. 205–210
  76. S. Oishi, T. Matsuba, M. Makino, A. Kai, in *Proc. of Interspeech*. Combining state-level spotting and posterior-based acoustic match for improved query-by-example spoken term detection (ISCA, France, 2016), pp. 740–744
  77. M. Obara, K. Kojima, K. Tanaka, S.-W. Lee, Y. Itoh, in *Proc. of Interspeech*. Rescoring by combination of posteriorgram score and subword-matching score for use in Query-by-Example (ISCA, France, 2016), pp. 1918–1922
  78. B. Taras, C. Nadeu, Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion. *EURASIP J. Audio Speech Music. Process.* **2011**(1), 1–10 (2011)
  79. M. Zelenák, H. Schulz, J. Hernando, Speaker diarization of broadcast news in Albayzin 2010 evaluation campaign. *EURASIP J. Audio Speech Music. Process.* **2012**(19), 1–9 (2012)
  80. L. J. Rodríguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, G. Bordel, in *Proc. of Interspeech*. The Albayzin 2010 Language Recognition Evaluation (ISCA, France, 2011), pp. 1529–1532
  81. 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)
  82. 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)
  83. 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)
  84. D. Castán, D. Tavaréz, P. Lopez-Otero, J. Franco-Pedroso, H. Delgado, E. Navas, L. Docio-Fernández, D. Ramos, J. Serrano, A. Ortega, E. Lleida, Albayzin-2014 evaluation: audio segmentation and classification in broadcast news domains. *EURASIP J. Audio Speech Music. Process.* **2015**(33), 1–9 (2015)
  85. F. Méndez, L. Docio, M. Arza, F. Campillo, in *Proc. of FALA*. The Albayzin 2010 text-to-speech evaluation (ISCA, France, 2010), pp. 317–340
  86. J. Tejedor, D. T. Toledano, P. Lopez-Otero, L. Docio-Fernandez, L. Serrano, I. Harnaez, 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)
  87. J. Tejedor, D. T. Toledano, P. Lopez-Otero, L. Docio-Fernandez, J. Proença, F. Perdigão, 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)
  88. J. Billa, K. W. Ma, J. W. McDonough, Zavalagkos, D. R. Miller, K. N. Ross, A. El-Jaroudi, in *Proc. of Eurospeech*. Multilingual speech recognition: the 1996 Byblos callhome system (ISCA, France, 1997), pp. 363–366
  89. H. Cuayahuitl, B. Serridge, in *Proc. of MICAI*. Out-of-vocabulary word modeling and rejection for spanish keyword spotting systems (Springer, Germany, 2002), pp. 156–165
  90. M. Killer, S. Stuker, T. Schultz, in *Proc. of Eurospeech*. Grapheme based speech recognition (ISCA, France, 2003), pp. 3141–3144
  91. J. Tejedor, *Contributions to keyword spotting and spoken term detection for information retrieval in audio mining*. PhD thesis. (Universidad Autónoma de Madrid, Madrid, 2009)
  92. 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 *Proc. of ICASSP*. Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models (IEEE, USA, 2010), pp. 4334–4337
  93. 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)
  94. J. Li, X. Wang, B. Xu, in *Proc. of Interspeech*. An empirical study of multilingual and low-resource spoken term detection using deep neural networks (ISCA, France, 2014), pp. 1747–1751

95. M. Hazewinkel, *Student test*. (Kluwer Academic, Denmark, 1994)
96. NIST, The spoken term detection (STD) 2006 Evaluation Plan. <https://catalog.ldc.upenn.edu/docs/LDC2011S02/std06-evalplan-v10.pdf>. Accessed Apr 2019
97. J. G. Fiscus, J. Ajot, J. S. Garofolo, G. Doddington, in *Proc. of SSSC*. Results of the 2006 spoken term detection evaluation (ACM, USA, 2007), pp. 45–50
98. A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, in *Proc. of Eurospeech*. The DET curve in assessment of detection task performance (ISCA, France, 1997), pp. 1895–1898
99. NIST, Evaluation Toolkit (STDEval) Software. <https://www.nist.gov/itl/iad/mig/tools>. Accessed Apr 2019
100. I. T. Union, 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 Apr 2019
101. E. Lleida, A. Ortega, A. Miguel, V. Bazán, C. Pérez, M. Zotano, A. de Prada, *RTVE2018 database description*. Vivolab and Corporación Radiotelevisión Española, Zaragoza. <http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf>. Accessed Apr 2019
102. 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)
103. N. Rajput, F. Metze, in *Proc. of MediaEval*. Spoken web search (CEUR, Germany, 2011), pp. 1–2
104. F. Metze, E. Barnard, M. Davel, C. van Heerden, X. Anguera, G. Gravier, N. Rajput, in *Proc. of MediaEval*. The spoken web search task (CEUR, Germany, 2012), pp. 41–42
105. X. Anguera, F. Metze, A. Buzo, I. Szöke, L. J. Rodríguez-Fuentes, in *Proc. of MediaEval*. The spoken web search task (CEUR, Germany, 2013), pp. 921–922
106. X. Anguera, L. J. Rodríguez-Fuentes, I. Szöke, A. Buzo, F. Metze, in *Proc. of MediaEval*. Query by Example Search on Speech at MediaEval 2014 (CEUR, Germany, 2014), pp. 351–352
107. I. Szöke, L. J. Rodríguez-Fuentes, A. Buzo, X. Anguera, F. Metze, J. Proenca, M. Lojka, X. Xiong, in *Proc. of MediaEval*. Query by Example Search on Speech at MediaEval 2015 (CEUR, Germany, 2015), pp. 81–82
108. T. Akiba, H. Nishizaki, H. Nanjo, G. J. F. Jones, in *Proc. of NTCIR-11*. Overview of the NTCIR-11 spokenquery&doc task (Japan Society for Promotion of Science, Japan, 2014), pp. 1–15
109. T. Akiba, H. Nishizaki, H. Nanjo, G. J. F. Jones, in *Proc. of NTCIR-12*. Overview of the NTCIR-12 spokenquery&doc-2 (Japan Society for Promotion of Science, Japan, 2016), pp. 1–13
110. P. Schwarz, *Phoneme recognition based on long temporal context*. PhD thesis. (FIT, BUT, Brno, Czech Republic, 2008)
111. A. Varona, M. Penagarikano, L. J. Rodríguez-Fuentes, G. Bordel, in *Proc. of Interspeech*. On the use of lattices of time-synchronous cross-decoder phone co-occurrences in a SVM-phonotactic language recognition system (ISCA, France, 2011), pp. 2901–2904
112. F. Eyben, M. Wollmer, B. Schuller, in *Proc. of ACM Multimedia (MM)*. OpenSMILE - the Munich versatile and fast open-source audio feature extractor (ACM, USA, 2010), pp. 1459–1462
113. Y. Zhang, J. R. Glass, in *Proc. of ASRU*. Unsupervised spoken keyword spotting via segmental DTW on gaussian posteriorgrams (IEEE, USA, 2009), pp. 398–403
114. 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 *Proc. of ASRU*. The KALDI speech recognition toolkit (IEEE, USA, 2011)
115. M. Muller, *Information retrieval for music and motion*. (Springer, New York, 2007)
116. I. Szöke, M. Skacel, L. Burget, in *Proc. of MediaEval*. BUT QUESST 2014 system description (CEUR, Germany, 2014), pp. 621–622
117. J. Ponte, W. Croft, in *Proc. of ACM SIGIR*. A language modeling approach to information retrieval, (1998), pp. 275–281
118. J. Parapar, A. Freire, A. Barreiro, in *Proc. of ECIR*. Revisiting n-gram based models for retrieval in degraded large collections, (2009), pp. 680–684
119. E. Rodríguez-Banga, C. Garcia-Mateo, F. Méndez-Pazó, M. González-González, C. Magariños, in *Proc. of Iberspeech*. Cotovia: an open source TTS for Galician and Spanish, (2012), pp. 308–315
120. C. Manning, P. Raghavan, H. Schütze, *Introduction to information retrieval*. (Cambridge University Press, Cambridge, 2008)
121. A. Abad, L. J. Rodríguez-Fuentes, M. Peñagarikano, A. Varona, G. Bordel, in *Proc. of Interspeech*. On the calibration and fusion of heterogeneous spoken term detection systems, (2013), pp. 20–24
122. N. Brummer, D. van Leeuwen, in *Proc. of IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*. On calibration of language recognition scores (IEEE, USA, 2006), pp. 1–8
123. N. Brummer, E. de Villiers, *The BOSARIS Toolkit user guide: theory, algorithms and code for binary classifier score processing*. (Agnitio Labs. <https://sites.google.com/site/nikobrummer>. Accessed Apr 2019
124. J. Wiseman, Python interface to the WebRTC (<https://webrtc.org/>) voice activity detector (VAD). <https://github.com/wiseman/py-webrtcvad>. Accessed Apr 2019
125. A. Silnova, P. Matejka, O. Glembek, O. Plchot, O. Novotny, F. Grezl, P. Schwarz, L. Burget, J. H. Cernocky, in *Proc. of Odyssey*. BUT/Phonexia bottleneck feature Extractor (IEEE, USA, 2018), pp. 283–287
126. C. Cieri, D. Miller, K. Walker, in *Proc. of LREC*. The Fisher Corpus: a resource for the next generations of speech-to-text (ELRA, Belgium, 2004), pp. 69–71
127. Intelligence Advanced Research Projects Activity (IARPA), *Babel Program*. (Intelligence Advanced Research Projects Activity (IARPA). <https://www.iarpa.gov/index.php/research-programs/babel>. Accessed Apr 2019
128. L. J. Rodríguez-Fuentes, A. Varona, M. Penagarikano, G. Bordel, M. Diez, in *Proc. of ICASSP*. High-performance query-by-example spoken term detection on the SWS 2013 evaluation (IEEE, USA, 2014), pp. 7819–7823
129. A. Abad, L. J. Rodríguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, G. Bordel, in *Proc. of Interspeech*. On the calibration and fusion of heterogeneous spoken term detection systems (ISCA, France, 2013), pp. 20–24
130. C. Garcia-Mateo, J. Dieguez-Tirado, L. Docio-Fernandez, A. Cardenal-Lopez, in *Proc. of LREC*. Transcrigal: a bilingual system for automatic indexing of broadcast news (ELRA, Belgium, 2004), pp. 2061–2064
131. A. Moreno, L. Campillos, in *Proc. of Iberspeech*. MAVIR: a corpus of spontaneous formal speech in spanish and english (ISCA, France, 2004), pp. 224–230
132. A. Stolcke, in *Proc. of Interspeech*. SRILM - an extensible language modeling toolkit (ISCA, France, 2002), pp. 901–904
133. G. Chen, S. Khudanpur, D. Povey, J. Trmal, D. Yarowsky, O. Yilmaz, in *Proc. of ICASSP*. Quantifying the value of pronunciation lexicons for keyword search in low resource languages (IEEE, USA, 2013), pp. 8560–8564
134. V. T. Pham, N. F. Chen, S. Sivasdas, H. Xu, I.-F. Chen, C. Ni, E. S. Chng, H. Li, in *Proc. of SLT*. System and keyword dependent fusion for spoken term detection (IEEE, USA, 2014), pp. 430–435
135. D. Can, M. Saraclar, Lattice indexing for spoken term detection. *IEEE Trans. Audio Speech Lang. Process.* **19**(8), 2338–2347 (2011)
136. D. R. H. Miller, M. Kleber, C.-L. Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, H. Gish, in *Proc. of Interspeech*. Rapid and accurate spoken term detection (ISCA, France, 2007), pp. 314–317
137. G. Chen, O. Yilmaz, J. Trmal, D. Povey, S. Khudanpur, in *Proc. of ASRU*. Using proxies for OOV keywords in the keyword search task (IEEE, USA, 2013), pp. 416–421

## Publisher's Note

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

Submit your manuscript to a SpringerOpen® journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)