Tejedor *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2018) 2018:2
https://doi.org/10.1186/s13636-018-0125-9

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

ALBAYZIN Query-by-example Spoken Term Detection 2016 evaluation

Javier Tejedor[1*], Doroteo T. Toledano[2], Paula Lopez-Otero[3], Laura Docio-Fernandez[4], Jorge Proença[5], Fernando Perdigão[5], Fernando García-Granada[6], Emilio Sanchis[6], Anna Pompili[7] and Alberto Abad[7]

Abstract

Query-by-example Spoken Term Detection (QbE STD) aims to retrieve data from a speech repository given an acoustic (spoken) query containing the term of interest as the input. This paper presents the systems submitted to the ALBAYZIN QbE STD 2016 Evaluation held as a part of the ALBAYZIN 2016 Evaluation Campaign at the IberSPEECH 2016 conference. Special attention was given to the evaluation design so that a thorough post-analysis of the main results could be carried out. Two different Spanish speech databases, which cover different acoustic and language domains, were used in the evaluation: the MAVIR database, which consists of a set of talks from workshops, and the EPIC database, which consists of a set of European Parliament sessions in Spanish. We present the evaluation design, both databases, the evaluation metric, the systems submitted to the evaluation, the results, and a thorough analysis and discussion. Four different research groups participated in the evaluation, and a total of eight template matching-based systems were submitted. We compare the systems submitted to the evaluation and make an in-depth analysis based on some properties of the spoken queries, such as query length, single-word/multi-word queries, and in-language/out-of-language queries.

Keywords: Query-by-example Spoken Term Detection, International evaluation, Spanish, Search on spontaneous speech

1 Introduction

The huge amount of heterogeneous speech data stored in audio and audiovisual repositories makes it necessary to develop efficient methods for speech information retrieval. There are different speech information retrieval tasks, including spoken document retrieval (SDR), keyword spotting (KWS), spoken term detection (STD), and query-by-example spoken term detection (QbE STD).

Spoken term detection aims at finding individual words or sequences of words within audio archives. It is based on a text-based input, commonly the word/phone transcription of the search term. For this reason, STD is also called text-based STD. Query-by-example spoken term detection is similar, but is based on an acoustic (spoken) input. In QbE STD, we consider the scenario in which the user has found a segment of speech which contains terms of interest within a speech data repository, and their purpose is to find similar speech segments within that repository. The speech segment found is the query, and the system outputs other similar segments from the repository, which we will henceforth refer to as utterances. Alternatively, the query can be uttered by the user. This is a highly valuable task for blind people or devices that do not have a text-based input, and consequently, the query must be given in other format such as speech.

The STD systems are typically composed of three different stages: (1) the audio is decoded into word/subword lattices using an automatic speech recognition (ASR) subsystem trained for the target language (which makes the STD system language-dependent), (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

*Correspondence: javier.tejedornoguerales@ceu.es
[1]Escuela Politécnica Superior. Universidad San Pablo-CEU, CEU Universities, Campus de Montepríncipe, Madrid, Spain
Full list of author information is available at the end of the article

amounts of resources in the form of transcribed corpora to be built.

QbE STD has been mainly addressed from 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 query

These methods make use of the text-based STD technology. In order to do this, they need to transcribe the query into word/subword units. The errors produced in this transcription can lead to significant performance degradation. [1, 2] employ a Viterbi-based search on Hidden Markov Models (HMMs). [3–6] employ dynamic time warping (DTW) or variants of DTW, e.g., non-segmental dynamic time warping (NS-DTW) from phone recognition. [7–10] employ word and syllable speech recognizers. Hou et al. [11] employs a phone-based speech recognizer and a weight finite-state transducer (WFST)-based search. Vavrek et al. [12] uses multilingual phone-based speech recognition, from supervised and unsupervised acoustic models and sequential dynamic time warping for search.

1.2 Methods based on template matching of features

These methods extract a set of features from the query and the speech repository, and a search of these features produces the query detections. Regarding the features used for query/utterance representation, [5, 13–15] employ Gaussian posteriorgrams; [16] proposes an *i*-vector-based approach for feature extraction; [17] uses phone log-likelihood ratio-based features; [18] employs posteriorgrams derived from various unsupervised tokenizers, supervised tokenizers, and semi-supervised tokenizers; [19] employs posteriorgrams derived from a Gaussian mixture model (GMM) tokenizer, phoneme recognition, and acoustic segment modelling; [11, 15, 20–26] use phoneme posteriorgrams; [11, 27–29] employ bottleneck features; [30] employs posteriorgrams from non-parametric Bayesian models; [31] employs articulatory class-based posteriorgrams; [32] proposes an intrinsic spectral analysis; and [33] is based on the unsupervised segment-based bag of an acoustic words framework.

All these studies employ the standard DTW algorithm for query search, except for [13], which employs the NS-DTW algorithm, [15, 24, 25, 28, 30], which employ the subsequence DTW (S-DTW) algorithm, [14], which presents a variant of the S-DTW algorithm, and [26], which employs the segmental DTW algorithm.

These methods were found to outperform subword transcription-based techniques in QbE STD [34]. This approach can be employed effectively to build language-independent STD systems, since prior knowledge of the language involved in the speech data is not necessary.

1.3 Hybrid approach

A powerful way of enhancing performance relies on building hybrid (fused) systems that combine the two individual methods. [35–37] propose a logistic regression-based fusion of acoustic keyword spotting and DTW-based systems using language-dependent phoneme recognizers. [38–41] use a logistic regression-based fusion on DTW- and phone-based systems. Oishi et al. [42] uses a 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, and [43] adds a logistic regression-based approach for detection rescoring. Obara et al. [44] employs a syllable-based speech recognizer and dynamic programming at the triphone-state level to output detections and DNN posteriorgram-based rescoring.

1.4 Motivation and organization of this paper

The increasing interest from within the speech research community in speech information retrieval has allowed the successful organization of several international evaluations related to SDR [45, 46], STD [47, 48], and QbE STD [49, 50]. In 2012 and 2014, the first two QbE STD evaluations in Spanish were held in the context of the ALBAYZIN 2012 and 2014 evaluation campaigns. These campaigns are internationally open sets of evaluations supported by the Spanish Network of Speech Technologies (RTTH)[1] and the ISCA Special Interest Group on Iberian Languages (SIG-IL)[2], which have been held every 2 years since 2006. These evaluation campaigns provide an objective mechanism for the comparison of different systems and the promotion of research into different speech technologies such as audio segmentation [51], speaker diarization [52], language recognition [53], spoken term detection [54], query-by-example spoken term detection [55, 56], and speech synthesis [57].

The Spanish language is widespread throughout the world, and significant research has been conducted into it for ASR [58–60], KWS [61, 62], and STD [62–64]. This, combined with the success of the ALBAYZIN QbE STD evaluations held in 2012 and 2014, have encouraged us to organize a new QbE STD evaluation for the 2016 ALBAYZIN evaluation campaign which aims to evaluate the progress in this technology in Spanish. Compared with the previous evaluations, the third ALBAYZIN QbE STD evaluation incorporated stricter rules regarding the evaluation queries, e.g., in-vocabulary (INV) vs. out-of-vocabulary (OOV) queries, and employs two different databases to cover different acoustic conditions and topics to provide a more comprehensive evaluation. In addition,

all the queries and the database employed in the QbE STD evaluation held in 2014 are kept, thus enabling a comparison between the systems submitted to both evaluations on the common set of queries.

The remainder of the paper is organized as follows: The following section presents a description of the QbE STD evaluation. Section 3 presents the different systems submitted to the evaluation. The results and discussion are then presented, and the paper is concluded in the final section.

2 ALBAYZIN QbE STD 2016 evaluation

2.1 Evaluation description

The ALBAYZIN QbE STD 2016 evaluation involves searching for audio content within audio content using an audio content query. The input to the system is an acoustic example per query; therefore, prior knowledge of the correct word/subword transcription corresponding to each query is not available. The target participants are the research groups or companies working on speech indexing, speech retrieval, and speech recognition.

The evaluation consists of searching a development query list within development speech data, and searching two different test query lists within two different sets of test speech data (MAVIR and EPIC databases, which will be explained later). The evaluation result ranking is based on the system performance when searching the query terms within the test speech data corresponding to the MAVIR database. Any kind of data, except for the MAVIR test data and the EPIC data, can be used by the participants for system training and development. The systems could be fine-tuned for each of the two databases individually. To facilitate the system construction, the participants were provided with MAVIR data, which can only be used as defined by the training, development, and test subsets.

This evaluation defines two different sets of queries for each database: the in-vocabulary query set and the out-of-vocabulary query set. The OOV query set was defined to simulate the out-of-vocabulary words of a Large Vocabulary Continuous Speech Recognition (LVCSR) system. If the participants employed an LVCSR system for processing the audio, these OOV queries must be removed from the system dictionary. Therefore, other methods must be used for searching the OOV queries. Conversely, the INV queries can appear in the dictionary of the LVCSR system.

The evaluation participants could submit a primary system and up to two contrastive systems. No manual intervention was allowed to generate the final output file, and hence, all the systems had to be fully automatic. Listening to the test data, or any other human interaction with the test data, was forbidden before all the evaluation results had been sent to the participants. The standard XML-based format corresponding to the National Institute of Standards and Technology (NIST) STD evaluation tool [65] was used to build the system output file.

The participants were given approximately 3 months to construct the system. The training and development data were released by the end of June 2016. The test data were released at the beginning of September 2016. The final system submission was due by mid-October 2016. The evaluation results were discussed at the IberSPEECH 2016 conference at the end of November 2016.

2.2 Evaluation metric

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 a *false alarm*. If an actual occurrence is not detected, this is called a *miss*. The actual term-weighted value (ATWV) proposed by NIST [65] was used as the main metric for the evaluation. This metric integrates the hit rate and the false alarm rate of each query into a single metric and is then averaged over all the queries:

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

where Δ 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 β is a weight factor set at 999.9, as in the ATWV proposed by NIST [66]. This weight factor causes an emphasis to be placed on recall compared to the precision in the ratio 10:1.

The ATWV represents the term-weighted value (TWV) for the threshold set by the system (usually tuned on development data). An additional metric, called maximum term-weighted value (MTWV) [65], can also be used to evaluate the performance of a QbE STD system. The MTWV is the ATWV the system would obtain with the optimum threshold. The MTWV results are presented to evaluate threshold selection.

In addition to the ATWV and the MTWV, NIST also proposed a detection error trade-off (DET) curve [67] to evaluate the system performance at various miss/FA ratios. Although the DET curves were not used for the evaluation, they are also presented in this paper for a comparison of the systems.

The NIST STD evaluation tool [68] was employed to compute the MTWV, the ATWV, and the DET curves.

2.3 Database

Two different databases that comprise different acoustic conditions and domains were employed for the evaluation. For comparison, the same MAVIR database employed in the ALBAYZIN QbE STD evaluation held in 2014 was

used. The second database was the EPIC database distributed by ELRA[3]. For the MAVIR database, three separate datasets, i.e., training, development, and test, were given to the participants. For the EPIC database, only the test data were provided. The MAVIR and EPIC data could only be used for the intended purpose of the corresponding subset (training, development, and test). The use of two different domains was permitted to compare the system performance across the two different domains and enabled the examination of the performance degradation of the systems depending on the nature of the speech data, the acoustic conditions, the training/development and testing mismatch, and the over-fitting issues.

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

The MAVIR Spanish data consist of spontaneous speech files, each containing different speakers, amounting to approximately 7 h of speech. These data were further divided for the purpose of this evaluation into training, development, and test sets. The data were also manually annotated in an orthographic form, but the timestamps were only set for the phrase boundaries. To prepare the data for the evaluation, the organizers manually added the timestamps for the approximately 1600 occurrences of the spoken terms used in the development and test evaluation sets. The training data were made available to the participants and included the orthographic transcription and the timestamps for the phrase boundaries[5].

The MAVIR speech data were originally recorded in several audio formats, e.g., pulse code modulation (PCM) mono and stereo, MP3, 22.05 KHz, and 48 KHz. The data were converted to PCM, 16 KHz, single channel, 16 bits per sample using the SoX tool[6]. Except for one, all the recordings were made with the same equipment, a Digital TASCAM DAT model DA-P1. Different microphones were used for the different recordings. In most cases, they were tabletop or floor standing microphones, but in one case, a lavalier microphone was used. The distance from the mouth of the speaker to the microphone varied and was not particularly controlled but in most cases was less than 50 cm. The recordings contain spontaneous speech from the MAVIR workshops in a real setting. The recordings were made in large conference rooms with capacity of over a hundred people, and a large number of people were in the conference room. This poses additional challenges including background noise, in particular 'babble noise' and reverberation. The realistic settings and the different nature of the spontaneous speech in this database made it appealing and challenging enough for the evaluation. Table 1 includes some database features such as the division in training, development, test data of the speech files, the number of word occurrences, the file duration, and the p.563 Mean Opinion Score (MOS) [69] which gives and indication of the quality of each speech file. The p.563 standard estimates the quality of the human voice without a reference signal, for which no reference signal is necessary. The MOS values are in the range of 1–5, 1 representing the worst quality and 5 the best [69].

The EPIC database comprises speeches from the European Parliament recorded in 2004 in English, Spanish, and Italian, together with their corresponding simultaneous translations into other languages. Only the original Spanish speeches, which consist of more than 1.5 h of clean speech, were used for the evaluation as a test set.

Table 1 Summary of MAVIR database

File ID	Data	No. of word occ.	Dur. (min)	No. of spk.	p.563 ave. MOS
Mavir-02	train	13432	74.51	7 (7 ma.)	2.69
Mavir-03	dev	6681	38.18	2 (1 ma. 1 fe.)	2.83
Mavir-06	train	4332	29.15	3 (2 ma. 1 fe.)	2.89
Mavir-07	dev	3831	21.78	2 (2 ma.)	3.26
Mavir-08	train	3356	18.90	1 (1 ma.)	3.13
Mavir-09	train	11179	70.05	1 (1 ma.)	2.39
Mavir-12	train	11168	67.66	1 (1 ma.)	2.32
Mavir-04	test	9310	57.36	4 (3 ma. 1 fe.)	2.85
Mavir-11	test	3130	20.33	1 (1 ma.)	2.46
Mavir-13	test	7837	43.61	1 (1 ma.)	2.48
ALL	train	43467	260.27	13 (12 ma. 1 fe.)	2.60
ALL	dev	10512	59.96	4 (3 ma. 1 fe.)	2.96
ALL	test	20277	121.3	6 (5 ma. 1 fe.)	2.67

'MOS' stands for Mean Opinion Score, as estimated using the ITU-T p.563 standard, *train* training, *dev* development, *occ.* occurrences, *dur.* duration, *spk.* speakers, *ma.* male, *fe.* female, *ave.* average

To evaluate the systems submitted to the evaluation, the organizers manually added the timestamps for the approximately 1100 occurrences of the spoken terms used in the test set.

The original speeches in the EPIC database were recorded as video files stored in a .mpeg1 format. Therefore, the original Spanish speeches were extracted from the corresponding Spanish video files, and converted to PCM, 16 KHz, single channel, 16 bits per sample, using the ffmpeg tool[7]. Table 2 includes the Spanish EPIC database with the same database features presented in Table 1.

2.3.1 Query list selection

All the queries selected 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) queries, out-of-language (OOL) queries, single-word and multi-word queries, in-vocabulary and out-of-vocabulary queries, and queries of different lengths. A query may not have any occurrence or may appear once or more in the speech data. Table 3 includes some features of the development and test query lists 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, together with the number of occurrences of each set in the corresponding speech data. It must be noted that a multi-word query was considered OOV in cases where any of the words that formed the term of the query were OOV. The test EPIC query list only contained *easy* terms, i.e., no OOL and multi-word queries were included, because this corpus was aimed at evaluating the systems submitted to the evaluation in a different domain.

2.4 Comparison with other QbE STD evaluations

The evaluations that are most similar to the ALBAYZIN QbE STD are the MediaEval 2011 [70], 2012 [71], and 2013 [49] Spoken Web Search 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 by the systems, and only the audio files that contained the query must be retrieved [46]. In 2015, the QUESST was similar to that of 2014, but the systems had to provide a score per query and

Table 2 Summary of EPIC database

File ID	No. of word occ.	Dur. (min)	No. of spk.	p.563 ave. MOS
10-02-04-m-058-org-es	280	2.47	1 fe.	3.71
10-02-04-m-074-org-es	3189	25.2	1 ma.	2.79
11-02-04-m-017-org-es	532	3.47	1 fe.	3.70
11-02-04-m-022-org-es	896	5.08	1 ma.	2.76
11-02-04-m-032-org-es	726	3.37	1 ma.	3.12
11-02-04-m-035-org-es	535	3.12	1 ma.	3.44
11-02-04-m-041-org-es	92	0.78	1 ma.	3.00
11-02-04-m-054-org-es	199	1.70	1 ma.	3.12
12-02-04-m-010-org-es	344	2.38	1 ma.	3.18
12-02-04-m-028-org-es	78	0.45	1 ma.	1.66
12-02-04-m-038-org-es	285	2.17	1 ma.	3.31
25-02-04-p-024-org-es	1205	8.50	1 fe.	3.92
25-02-04-p-027-org-es	353	2.23	1 ma.	3.67
25-02-04-p-030-org-es	523	3.18	1 fe.	3.79
25-02-04-p-034-org-es	353	2.23	1 fe.	3.78
25-02-04-p-037-org-es	492	2.93	1 fe.	3.67
25-02-04-p-043-org-es	1705	12.27	1 ma.	3.32
25-02-04-p-047-org-es	922	5.82	1 ma.	3.39
25-02-04-p-072-org-es	278	1.90	1 fe.	4.27
25-02-04-p-081-org-es	1270	8.07	1 ma.	3.20
25-02-04-p-096-org-es	211	1.27	1 ma.	3.41
ALL	14468	98.58	21 (14 ma. 7 fe.)	3.20

'MOS' stands for Mean Opinion Score, as estimated using the ITU-T p.563 standard, *train* training, *dev* development, *occ.* occurrences, *dur.* duration, *spk.* for speakers, *ma.* male, *fe.* female, *ave.* for average

Table 3 Statistics of the development and the test query lists for the MAVIR and the EPIC databases

Query list	dev	test-MAVIR	test-EPIC
#INL queries (occ.)	96 (386)	99 (1163)	95 (1152)
#OOL queries (occ.)	6 (39)	7 (29)	0 (0)
#SINGLE queries (occ.)	93 (407)	100 (1180)	95 (1152)
#MULTI queries (occ.)	9 (18)	6 (12)	0 (0)
#INV queries (occ.)	83 (296)	94 (979)	78 (917)
#OOV queries (occ.)	19 (129)	12 (213)	17 (235)

The length of the development queries varies between 5 and 21 phonemes. The length of the MAVIR test queries varies between 5 and 16 phonemes. The length of the EPIC test queries varies between 5 and 15 phonemes. *dev* development, *INL* in-language queries, *OOL* out-of-language queries, *SINGLE* single-word queries, *MULTI* multi-word queries, *INV* in-vocabulary queries, *OOV* out-of-vocabulary queries, *occ.* occurrences

utterance [72]. 2016 was the last year that the search-on-speech task was included in MediaEval, by means of the *zero-cost speech recognition task*. This consisted of building LVCSR systems from low resources [73]. The task in the MediaEval 2011, 2012, and 2013 Spoken Web Search and the ALBAYZIN evaluations was the same, i.e., searching speech content from speech queries, but they differed in several aspects. This makes it difficult to compare the results obtained in the ALBAYZIN QbE STD evaluation to the previous MediaEval evaluations.

The most important difference is the nature of the audio content used for the evaluations. In the MediaEval evaluations, the speech was typically telephone speech, either conversational or read and elicited speech, or speech recorded with in-room microphones. In the ALBAYZIN 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. The microphones, the conference rooms, and the recording conditions changed from one recording to another. The microphones were not close-talking microphones but were mainly tabletop or floor standing microphones.

In addition, the MediaEval evaluations dealt with Indian- and African-derived languages, as well as Albanian, Basque, Czech, non-native English, Romanian, and Slovak languages, while the ALBAYZIN evaluations deal only with Spanish.

In addition to the MediaEval evaluations, a new round of QbE STD evaluations was organized with the NTCIR-11 [74] and NTCIR-12 [75] conferences. The data used in these evaluations contained spontaneous speech in Japanese provided by the National Institute for Japanese language and spontaneous speech which was recorded during seven editions of the Spoken Document Processing Workshop. As additional information, these evaluations provided participants with the results of a voice activity detection system on the input speech data, the manual transcription of the speech data, and the output of an LVCSR system. Although the ALBAYZIN QbE STD evaluation could be considered to be similar in terms of speech nature to the NTCIR QbE STD evaluations, i.e., the speech was recorded in real workshops, the ALBAYZIN evaluations make use of other languages and define disjointed development and test query lists to measure the generalization capability of the systems.

Table 4 summarizes the main characteristics of the MediaEval QbE STD evaluations, the NTCIR-11 and NTCIR-12 QbE STD evaluations, the previous ALBAYZIN QbE STD evaluations, and the ALBAYZIN QbE STD 2016 evaluation.

Table 4 Comparison of the different QbE STD evaluations: Albanian (ALB), Basque (BAS), Czech (CZE), non-native English (NN-ENG), Isixhosa (ISIX), Isizulu (ISIZ), Romanian (ROM), Sepedi (SEP), Setswana (SET), and Slovak (SLO)

Evaluation	Language/s	Type of speech	No. of 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

Tel. telephone, *mic.* microphone, *dev.* development, *ATWV* actual term weighted value, C_{nxe} normalized cross entropy cost, *MAP* mean average precision

3 Systems

Eight different systems were submitted to the ALBAYZIN QbE STD 2016 evaluation from four different research groups (see Table 5). Some were submitted in time for the evaluation, and some were submitted as post-evaluation systems and so were not included in the competition. All were based on a feature representation of the queries and the utterances and a DTW-based search to hypothesize detections. In addition, a text-based STD system was also presented to compare performance when using written and acoustic (spoken) queries.

3.1 A-GTM-UVigo-Three feature+DTW-based fusion QbE STD system (A-GTM-UVigo-3-fea+DTW fusion)

The architecture of this system is shown in Fig. 1; it consists of a fusion of three different DTW-based QbE STD systems that employ different approaches for feature extraction.

3.1.1 Feature extraction

Given a query Q with n frames (and equivalently, an utterance U with m frames), three speech representations that result in a set $\mathrm{Q} = \{\mathrm{q}_1, \ldots, \mathrm{q}_n\}$ of n vectors of dimension D (and equivalently, a set of $\mathrm{U} = \{\mathrm{u}_1, \ldots, \mathrm{u}_m\}$ of m vectors of dimension D) are based on:

- Phoneme posteriorgram + phoneme unit selection: This speech representation relies on phoneme posteriorgrams [34]. Given a query/utterance and a phoneme recognizer with P phonetic units, the posterior probability of each phonetic unit is computed for each frame, leading to a set of vectors of dimension P that represent the probability of each phonetic unit at every frame. To construct a wide-coverage language-independent QbE STD system, the Czech, English, Hungarian, and Russian phoneme recognizers developed by the Brno University of Technology (BUT) [76] are used to obtain the phoneme posteriorgrams; in these decoders, each phonetic unit has three different states and a posterior probability is an output for each of them, so they are combined to obtain one posterior probability for each unit [17]. After obtaining the posteriors, Gaussian softening is applied to obtain Gaussian-distributed probabilities [77]. Then, the phoneme unit selection strategy described in [25] is applied.
- Acoustic features + feature selection: Aiming to obtain as much information as possible from the speech signals, a large set of features, summarized in Table 6, are used to represent the queries and utterances; these features, obtained using the OpenSMILE feature extraction toolkit [78], are extracted every 10 ms using a 25-ms window, except for the F0, probability of voicing, jitter, shimmer, and harmonics-to-noise ratio (HNR), where a 60 ms window is used due to its best performance in preliminary work. Finally, the feature selection technique described in [79] is applied to obtain the most discriminative features.
- Gaussian posteriorgrams: The Gaussian posteriorgrams [80] are used to represent the queries and the utterances. Given a GMM with G Gaussians, the posterior probability of each Gaussian is computed for each time frame, leading to a set of vectors of dimension G that represent the probability of each Gaussian at every time instant. In this system, 19 Mel-frequency Cepstral Coefficients (MFCCs) are extracted from the acoustic signals, accompanied by their energy, delta, and double delta coefficients due to their best performance in previous work. The feature extraction and the Gaussian posteriorgram computation are carried out using the Kaldi toolkit [81].

3.1.2 Search

The search stage uses the S-DTW algorithm [82], which is a variant of the standard DTW. For the S-DTW, a cost matrix $\mathrm{M} \in \Re^{\mathrm{n}\times\mathrm{m}}$ must first be defined, in which the rows and the columns correspond to the query and the utterance frames, respectively:

$$\mathrm{M}_{i,j} = \begin{cases} c\left(\mathrm{q}_i, \mathrm{u}_j\right) & \text{if} \quad i = 0 \\ c\left(\mathrm{q}_i, \mathrm{u}_j\right) + \mathrm{M}_{i-1,0} & \text{if} \quad i > 0,\ j = 0 \\ c\left(\mathrm{q}_i, \mathrm{u}_j\right) + \mathrm{M}^*(i,j) & \text{else,} \end{cases} \tag{2}$$

Table 5 Participants in the ALBAYZIN QbE STD 2016 evaluation along with the submitted systems

Team ID	Research institution	Systems
GTM-UVigo	AtlantTIC Research Center Universidad de Vigo, Spain	A-GTM-UVigo-3-fea+DTW fusion (in-time)
L2F	L^2F Spoken Language Systems Lab, INESC-ID University of Lisbon, Portugal	B-L2F-4-pllr fea+DTW fusion (post-evaluation) C-L2F-4-likel fea+DTW fusion (post-evaluation)
ELiRF-UPV	Natural Language Engineering and Pattern Recognition Universitat Politécnica de Valéncia, Spain	D-ELiRF-UPV-Post+DTW (in-time) E-ELiRF-UPV-Post+DTWNorm (in-time)
SPL-IT-UC	Instituto de Telecomunicações University of Coimbra, Portugal	F-SPL-IT-UC-4-phnrec+DTW fusion (in-time) G-SPL-IT-UC-3-phnrec+DTW fusion (in-time) H-SPL-IT-UC-2-LIphnrec+DTW fusion (in-time)

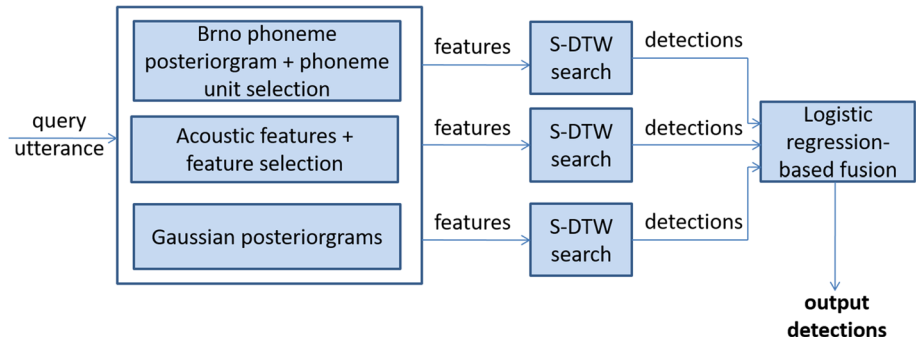

Fig. 1 A-GTM-UVigo-Three feature+DTW-based fusion QbE STD system architecture

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

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

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

Pearson's correlation coefficient r [83] is used as a cost function by mapping it into the interval [0,1] applying the following transformation:

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

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

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

The starting point of the path ending at b^*, namely a^*, is computed by backtracking, hence obtaining the best warping path $P(Q, U) = \{p_1, \ldots, p_k, \ldots, p_K\}$, where $p_k = (i_k, j_k)$, (i.e., the kth element of the path is formed by q_{i_k} and u_{j_k}, and K is the length of the 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 of 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 ∞ 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 [35], and then, z-norm is applied so that all the confidence scores of all the queries have the same distribution [37].

3.1.3 Fusion

Discriminative calibration and fusion are applied to combine the detections of the different systems obtained from the different feature extraction approaches [38]. The global minimum score produced by the systems for all the queries is used to hypothesize the missing confidence scores due to its good performance in previous work. The calibration and the fusion parameters are then estimated by logistic regression on the development data to obtain improved discriminative and well-calibrated likelihood ratios [84]. The calibration and the fusion training are performed using the Bosaris toolkit [85].

The fusion is carried out on the detections output by the S-DTW search from the phoneme posteriorgram + phoneme unit selection approach on the English phoneme decoder, the acoustic features + feature selection approach from a set of 90 relevant features, and the Gaussian posteriorgram approach with 128 Gaussians. This configuration proved to be the best on the development data.

3.2 B-L2F-Four phone log-likelihood ratio feature+DTW-based fusion QbE STD system (B-L2F-4-pllr fea+DTW fusion)

Four different QbE STD systems that employ DTW-based query detection and several phoneme recognizers are fused. The system architecture is shown in Fig. 2.

3.2.1 Speech segmentation

The set of utterances is pre-processed using the audio segmentation module presented in [86]. This performs speech/non-speech classification and speaker segmentation, as well as other tasks. The speech/non-speech segmentation is implemented using a multi-layer perceptron (MLP) based on perceptual linear prediction (PLP) features, followed by a finite state machine. This finite state machine smooths the input probabilities given by the MLP using a median filter over a small window. The smoothed signal is then thresholded and analysed using a time window ($t_{\min}$). The finite state machine consists of four possible states classified as *probable non-speech*, *non-speech*, *probable speech*, and *speech*. If the input audio signal has a probability of *speech* above a given threshold, the finite state machine is placed into the *probable speech*

Table 6 Acoustic features used in the A-GTM-UVigo-Three feature+DTW-based fusion system

Description	No. 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 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
MFCCs	16
MFCC filterbank	26
Line spectral pairs	8
Cepstral PLP 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
LPC formant frequencies and bandwidths	6
Formant frame intensity	1
First derivative	102
Total	204

RASTA log-RelAtive SpecTrAl, *MFCC* Mel-frequency cepstral coefficient, *PLP* perceptual linear prediction, *LPC* linear prediction coding

state. If, after a given time interval ($t_{\min}$), the average speech probability is above a given confidence value, the machine changes to the *speech* state. Otherwise, it goes to the *non-speech* state. The finite-state machine generates segment boundaries for the *non-speech* segments larger than the resolution of the median window. Additionally, the *non-speech* segments larger than $t_{\min}$ are discarded. The value of $t_{\min}$ and the threshold are chosen to maximize the non-speech detection in the work presented in [86] which aims to avoid the system processing the short silence segments included in large speech segments. With the speech segmentation module, a partition of each utterance into smaller segments is obtained. Only the resulting *speech* segments are given to the query search. This strategy offers two computational advantages: (1) Because the same query may occur multiple times in an utterance, a DTW-based search should proceed sequentially or iteratively over the whole utterance, storing the candidates found during the search, and initiating a new process with the remaining audio until a certain stopping criterion is met. By splitting the utterance into smaller segments, the search can be parallelized, allowing for different searches of the same query at the same time. (2) Because the segments classified as *non-speech* are discarded, the performance of the DTW algorithm benefits from an overall reduction in the search space. On the other hand, this strategy conveys at least two drawbacks that may affect the query detection: (1) The errors of the audio segmentation module can result in missing speech segments that may eventually prove to contain query terms that are lost. (2) It is assumed that only a single match per query can occur in a sub-segment, which may also introduce misses in the search.

3.2.2 Feature extraction

Two different approaches are employed for feature extraction which aim to obtain complementary information from the speech signals. The first employs the AUDIMUS phoneme recognizers for speech representation, and the second is based on the phoneme recognizers developed by the BUT [76].

The AUDIMUS phoneme recognizers are based on hybrid connectionist methods [87]. Four phoneme recognizers that exploit four different sets of acoustic models were used. These are trained in European Portuguese, Brazilian Portuguese, Spanish, and the American English languages. The acoustic models are based on MLPs that are part of the L^2F in-house hybrid connectionist ASR system called AUDIMUS [88, 89]. AUDIMUS combines four MLP outputs trained with various sets of features, as shown in Table 7. The language-dependent MLPs are trained using different amounts of annotated data. Each MLP is characterized by the input frame context size, i.e., 13 for PLP, PLP with log-RelAtive SpecTrAl (PLP-RASTA) and European Telecommunications Standards Institute (ETSI) features, and 28 for Modulation SpectroGram (MSG) features, the number of units of the two hidden layers (500), and the size of the output layer. Only monophone units are modelled, which results in 41-dimensional posterior vectors for English, 39-dimensional posterior vectors for Portuguese, 40-dimensional posterior vectors for Brazilian, and 30-dimensional posterior vectors for American English. These configurations are used due to their good performance in previous work. Finally, the frames for which the non-speech posterior

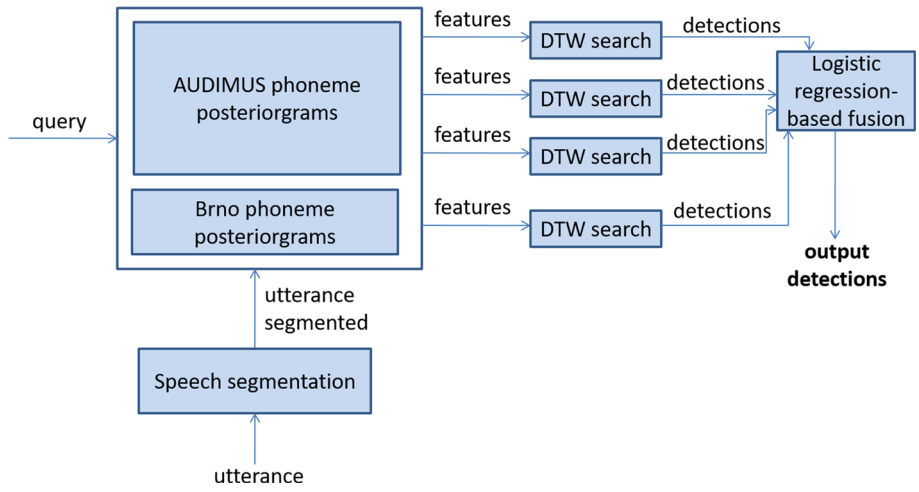

Fig. 2 B-L2F-Four phone log-likelihood ratio feature+DTW-based fusion QbE STD system architecture

probability is the highest are considered to be silence and discarded.

The phoneme recognizers for the Czech, Hungarian, and Russian languages developed by BUT [76] are also employed. These output phone-state level posterior probabilities and multiple non-speech units, which are reduced to single-state phone posterior probabilities, and a unique silence output unit. This results in 43-dimensional feature vectors for Czech, 59-dimensional feature vectors for Hungarian, and 50-dimensional feature vectors for Russian. The frames where the non-speech posterior probability is the highest are also discarded.

Finally, both the AUDIMUS and the BUT posterior feature vectors are converted to phone log-likelihood ratios (PLLR) as described in [90]. This representation proved to be very effective in spoken language recognition [91] and other similar tasks [92].

3.2.3 Search

Given two sequences of feature vectors corresponding to a query Q and an utterance U, the logarithm of the cosine distance is computed between each pair of vectors (Q[i], U[j]) to build a cost matrix as follows:

Table 7 Features used in the AUDIMUS decoders

Feature	No. of features
PLP	13 static + first derivative
PLP-RASTA	13 static + first derivative
MLG	28 static
Advanced front-end from ETSI	13 static + first and second derivatives

PLP perceptual linear prediction, PLP-RASTA PLP log-RelAtive SpecTrAl, MSG modulation SpectroGram, ETSI European Telecommunications Standards Institute

$$d(Q[i]\,,U[j]\,) = -log\frac{Q[i]\cdot U[j]}{|Q[i]\,|\cdot|U[j]\,|}. \tag{6}$$

The cost matrix is then normalized with respect to the utterance, such that the matrix values range from 0 to 1 [93]. The normalization is conducted as follows:

$$d_{\text{norm}}(Q[i]\,,U[j]\,) = \frac{d\left(Q[i]\,,U[j]\,\right) - d_{\min}(i)}{d_{\max}(i) - d_{\min}(i)}, \tag{7}$$

where $d_{\min}(i) = \min_{j=1,\ldots,n} d\left(Q[i]\,,U[j]\right)$ and $d_{\max}(i) = \max_{j=1,\ldots,n} d(Q[i]\,,U)$.

In this way, a perfect match would produce a quasi-diagonal sequence of zeros. The DTW search looks for the best alignment of the query and a partition of the normalized cost matrix corresponding to a *speech* segment. The algorithm uses three additional matrices to store the accumulated distance of the optimal partial warping path found (AD), the length of the path (L), and the path itself.

The best alignment of a query in an utterance is defined as the one that minimizes the average distance in a warping path of the normalized cost matrix. A warping path may start at any given frame of U, i.e., k_1, then traverses a region of U, which is optimally aligned to Q, and ends at frame k_2. The average distance in this warping path is computed as follows:

$$d_{\text{avg}}(Q,U) = AD[i,j]\,/L[i,j]\,.$$

The confidence score for each detection is computed as $1 - d_{\text{avg}}(Q,U)$, thus ranging from 0 to 1, where 1 represents a perfect match. The start time and the duration of each detection are obtained by retrieving the time offsets corresponding to the frames k_1 and k_2 in the filtered utterance. The detection results are filtered out to reduce the number of detections per query to a fixed amount of

hypothesis. Different values, ranging from 50 to 500, are experimented with to empirically determine the threshold, with the value of 100 detections per hour with the best performance observed on the development data.

3.2.4 Fusion

The output detections from the Brazilian Portuguese, Spanish, and European Portuguese AUDIMUS phoneme recognizers, and the Czech BUT phoneme recognizer [76], are fused with the strategy presented in the three feature+DTW-based fusion QbE STD system. This configuration gave the best performance on the development data.

3.3 C-L2F-Four likelihood feature+DTW-based fusion QbE STD system (C-L2F-4-likel fea+DTW fusion)

This system is the same as the B-L2F-Four phone log-likelihood ratio feature+DTW-based fusion QbE STD system with the following modifications:

- The English phoneme recognizer developed by BUT [76] is added to the feature extraction module.
- The fusion is carried out on the detections provided by the Brazilian Portuguese, Spanish, and English AUDIMUS phoneme recognizers and the English phoneme recognizer from BUT.
- The feature extractor from the AUDIMUS and the BUT phoneme recognizers [76] outputs log likelihoods instead of PLLR features.
- The threshold in the search is set to 300 detections per hour. This value was tuned on the development data with the new configuration.

3.4 D-ELiRF-UPV-Posteriorgram+DTW-based QbE STD system (D-ELiRF-UPV-Post+DTW)

This system, whose architecture is shown in Fig. 3, is based on DTW search on phoneme posteriorgrams.

For feature extraction, the phoneme recognizers developed at BUT for Czech, English, Hungarian, and Russian languages [76] are employed to obtain a posteriorgram-based representation of the queries and the utterances. The English language is employed in the final system submitted because this gave the best performance on the development data.

For a search, the system employs the S-DTW algorithm explained above. However, instead of using the usual transition set with horizontal, vertical, and diagonal path movements, the horizontal and vertical transitions are modified so that the paths found must have a length between half and twice the length of the query, as shown in Fig. 4. These path movement modifications aim to augment the query detection rate. To do so, $\mathrm{M}^*(i,j)$ in the cost matrix is modified as follows:

$$\mathrm{M}^*(i,j) = \min\left(\mathrm{M}_{i-x,j-y}\right), \tag{8}$$

where x and y represent the allowed transitions.

Different cost functions such as the Kullback-Leibler divergence, the cosine distance, and the inner product were explored, but the cosine distance was finally employed because it provided the best performance on the development data. The confidence score assigned to each detection is based on the distance computed by the S-DTW algorithm.

3.5 E-ELiRF-UPV-Posteriorgram+DTW-based normalized QbE STD system (E-ELiRF-UPV-Post+DTWNorm)

This system is the same as the D-ELiRF-UPV-Posteriorgram+DTW-based QbE STD system with a single modification in the S-DTW algorithm. This modification relies on the fact that the S-DTW search considers the length of the paths, and hence, the cost matrix is modified as follows:

$$\mathrm{M}_{i,j} = c\left(\mathrm{q}_i, \mathrm{u}_j\right) + \mathrm{M}\left(i - x', j - y'\right), \tag{9}$$

so that:

$$\left(x', y'\right) = \arg\min_{(x,y)} \frac{\mathrm{M}(i-x, j-y) + c(\mathrm{q}_i, \mathrm{u}_j)}{\mathrm{L}(i-x, j-y) + 1}, \tag{10}$$

where $\mathrm{L}(i-x, j-y)$ is the length of the best path ending in (i,j). With this modification, the fact that two paths have similar distance values but differ in the length of their alignments is considered.

3.6 F-SPL-IT-UC-Four phoneme recognizer+DTW-based fusion QbE STD system (F-SPL-IT-UC-4-phnrec+DTW fusion)

This system, whose architecture is presented in Fig. 5, consists of fusion of four DTW-based search systems from different phoneme recognizers.

3.6.1 Feature extraction

State-level phone posterior probabilities are employed as features for the query and the utterance representation.

Fig. 3 D-ELiRF-UPV-Posteriorgram+DTW-based QbE STD system architecture

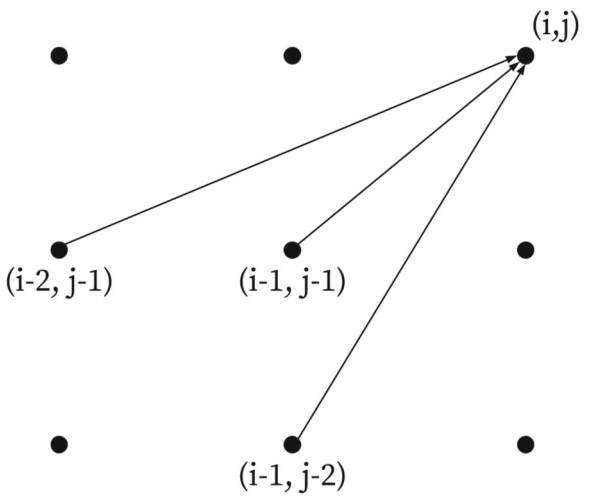

Fig. 4 Transitions allowed in the S-DTW search of the D-ELiRF-UPV-Posteriorgram+DTW-based QbE STD system

These are computed using the phoneme recognizer developed by BUT [76]. Three different phoneme recognizers are trained in Spanish, English, and European Portuguese. Although the queries mainly contain speech, a voice activity detector is employed. To do so, the frames for which the average of the posterior probability of silence and noise is higher than 0.5 were removed before applying the query search.

The Spanish recognizer was trained using the training data provided by the organizers. Because the file mavir02.wav presents a low-frequency noise, high-pass filtering with a cut-off frequency of 150 Hz, followed by spectral subtraction, is applied to this file before further processing. A phoneme dictionary is built using *g2p-seq2seq*[8] and a Spanish dictionary from CMU[9]. The phoneme alignment of the speech data is carried out with the Kaldi speech recognition toolkit [81].

As in previous studies [22, 94], the English recognizer was trained using the training subsets of TIMIT and Resource Management databases.

The European Portuguese recognizer was trained using annotated broadcast news data and a dataset of command words and sentences, as carried out in previous studies [22, 94].

3.6.2 Search

The DTW algorithm is used for query detection from the state-level phone posterior probabilities that represent each query and utterance frame. The logarithm of the cosine distance, as in the B-L2F-Four phone log-likelihood ratio feature+DTW-based fusion QbE STD system, is employed as a distance metric between a query and an utterance frame to build a cost matrix.

The DTW search considers paths that start at the first frame of the query and at any frame of the utterance and move in unitary weighted jumps diagonally, vertically, or horizontally from the lowest accumulated distance. The DTW search result corresponds to the accumulated distances (D_{acc}) at the last frame of the query, for every frame of the utterance. The information regarding

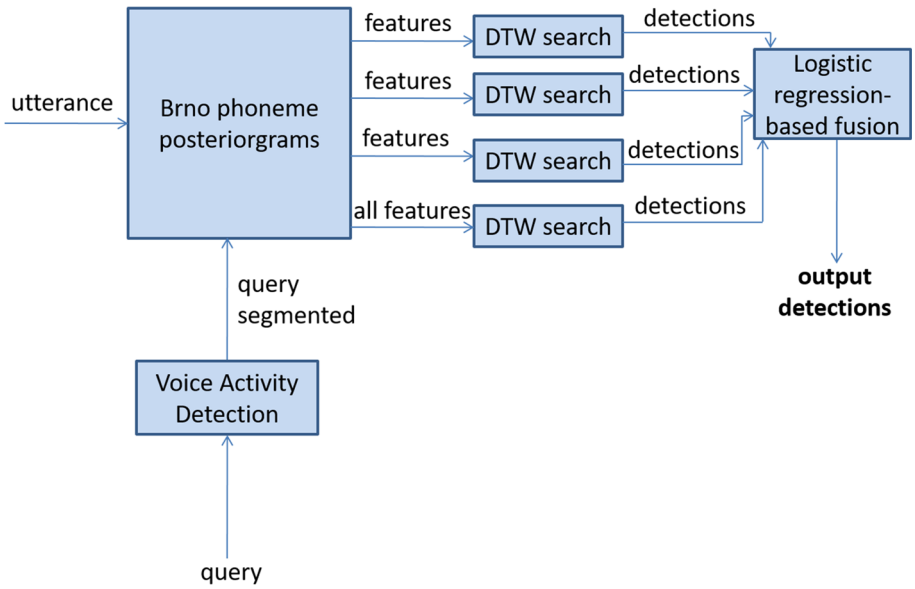

Fig. 5 F-SPL-IT-UC-Four phoneme recognizer+DTW-based fusion QbE STD system architecture

the start frame of the path, the ending frame of the path, and the number of diagonal, horizontal, and vertical movements is stored. The DTW search is carried out for Spanish, English, and European Portuguese languages individually. An additional DTW search based on averaging all the cost matrices given by the three languages is conducted, as in [18].

Finally, the accumulated distances are normalized according to the following equation:

$$D_{\text{norm}} = \frac{D_{\text{acc}}}{N_D + \frac{1}{2}(N_V + N_H)}, \tag{11}$$

where D_{acc} is the accumulated distance, and N_D, N_V, and N_H are the numbers of diagonal, vertical, and horizontal path movements, respectively. A confidence score is assigned to each detection by changing the sign of D_{norm}, i.e., score = $-D_{\text{norm}}$.

To select the candidate hits on the final normalized path distances, the system employs two limits for peak picking. The first is a hard limit of a maximum number of peaks, which implies an average of 1 peak per 20 s of audio. The second is a threshold where only the peaks above the 90% quantile of values above the mean plus standard deviation are selected. This guarantees that a small number of peaks is always chosen. Additionally, the peaks must be separated by a distance which is at least equal to the query length. The duration of the candidate hits in the utterance is also limited to between 0.5 and 1.9 times the size of the query. These figures were optimized on the development data.

3.6.3 Fusion

The next step is to normalize the confidence scores per query, for which z-norm is applied to each query score (q-norm). At this stage, there are four outputs from the four DTW search processes, i.e., the three phoneme recognizers and the average cost matrix. The fusion scheme is similar to that presented in [38]. Firstly, all the candidate hits are aligned (expanding the start and the end times), and a default score per sub-system for the candidate hits that are not found in all the sub-systems is assigned. This default score, which is equal to zero due to the q-norm, is the mean confidence score of that subsystem since this outperforms all other strategies such as the minimum score per query. All the candidate hits are considered, since this performs better than limiting the detections to those candidate hits found on more than one sub-system. Finally, the sub-system fusion is carried out by logistic regression with the Bosaris toolkit [85] to obtain improved discriminative and well-calibrated likelihood ratios [84]. The logistic regression is trained with the development data.

3.7 G-SPL-IT-UC-Three phoneme recognizer+DTW-based fusion QbE STD system (G-SPL-IT-UC-3-phnrec+DTW fusion)

This system is the same as the F-SPL-IT-UC-Four phoneme recognizer+DTW-based fusion QbE STD system except that the detections of the sub-system that employs the DTW-search on the average cost matrix are removed in the fusion strategy. This aims to evaluate the QbE STD system performance based on the individual languages.

3.8 H-SPL-IT-UC-Two language-independent phoneme recognizer+DTW-based fusion QbE STD system (H-SPL-IT-UC-2-LIphnrec+DTW fusion)

This system is the same as the F-SPL-IT-UC-Four phoneme recognizer+DTW-based fusion QbE STD system except that only the detections of the systems which employ the English and the Portuguese phoneme recognizers are fused. This aims to evaluate the QbE STD system performance using a language-independent setup.

3.9 I-Text-based STD system

This system was employed for comparison purposes with the QbE STD systems submitted to the evaluation. It was not submitted by any participant, nor did it compete in the evaluation. Because this system employs the correct transcription of the queries for the search, the system does not follow the rules of the evaluation. Therefore, this system simulates a scenario in which the queries are perfectly decoded by an ideal ASR subsystem.

The text-based STD system consists of the combination of a word-based STD system to detect the INV words and a phone-based STD system to detect the OOV words. Therefore, the correct word transcription of each query is used for the word-based STD system, and the correct phone transcription of each query is used for the phone-based STD system. Both systems are described below.

3.9.1 Word-based STD system

The ASR subsystem is based on the Kaldi open-source toolkit [81] and employs the DNN-based acoustic models. Specifically, a DNN-based context-dependent speech recognizer is trained following the DNN training approach presented in [95]. Forty-dimensional MFCCs, which are augmented with three pitch- and voicing-related features [96] and appended with their delta and double delta coefficients, are firstly extracted for each speech frame. The DNN has 6 hidden layers with 2048 neurons each. Each speech frame is spliced across ± 5 frames to produce 1419-dimensional vectors that are the input into the first layer. The output layer is a soft-max layer representing the log-posteriors of the context-dependent HMM states. The Kaldi LVCSR decoder generates word lattices [97] using these DNN-based acoustic models.

The data used for acoustic model (AM) training of this Kaldi-based LVCSR system have been extracted from the Spanish material in the 2006 TC-STAR automatic speech recognition evaluation campaign[10] and the Galician broadcast news database Transcrigal [98]. It must be noted that all the non-speech parts, as well as the speech parts corresponding to transcriptions with pronunciation errors, incomplete sentences, and short speech utterances, were discarded. This resulted in approximately 104.5 h of acoustic training material.

The language model (LM) of the LVCSR system is constructed using a text database of 160 millions of word occurrences from several sources such as the transcriptions of European and Spanish Parliaments from the TC-STAR database, subtitles, books, newspapers, online courses, and the transcriptions of the development data provided by the organizers. Specifically, the LM is obtained by static interpolation of the trigram-based LMs which are trained using these different text databases. The LMs are built with the SRILM toolkit [99], with the Kneser-Ney discounting strategy. The final interpolated LM is obtained using the SRILM static n-gram interpolation functionality. The LM vocabulary size is limited to the most frequent 60,000 words, and for each evaluation data set, the OOV terms are removed from the LM. This word-based LVCSR system configuration was chosen due to its good performance in the STD task [100].

The STD subsystem integrates the Kaldi term detector [81, 101, 102] which searches for the input terms within the word lattices obtained in the previous step. These lattices are processed using the lattice indexing technique described in [103] so that the lattices of all the utterances in the search collection are converted from the individual WFSTs to a single generalized factor transducer structure in which the start-time, the end-time, and the lattice posterior probability of each word token are stored as three-dimensional costs. This factor transducer is an inverted index of all the word sequences seen in the lattices. Thus, given a list of terms, a simple finite-state machine is created such that it accepts each term and composes it with the factor transducer to obtain all the occurrences of the terms in the search collection. The Kaldi decision-maker conducts a YES/NO decision, for each detection, based on the term-specific threshold (TST) approach presented in [104]. Therefore, a detection is assigned the YES decision if:

$$p > \frac{N_{\text{true}}}{\frac{T}{\beta} + \frac{\beta-1}{\beta} N_{\text{true}}}, \tag{12}$$

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

3.9.2 *Phone-based STD system*

The OOV terms are handled with a phone-based STD system strategy. A phoneme sequence is first obtained from the 1-best word path of the word-based Kaldi LVCSR system presented above. Next, a reduction of the phoneme set is performed to combine the phonemes with high confusion, which aims to augment the term detection rate; specifically, the semivowels /j/ and /w/ are represented as the vowels /i/ and /u/, respectively, and the palatal n /η/ is represented as /n/. Then, the *tre-agrep* tool is employed to compute candidate hits so that the Levenshtein distance between each recognized phoneme sequence and the phoneme sequence corresponding to each term can be computed. An analysis of the proposed strategy suggests that those candidate hits whose Levenshtein distance was equal to 0 were, in general, correct hits. The candidate hits with Levenshtein distance equal to 1 were found to be false alarms, although many hits were also found; since no specific criterion to assign a confidence score is implemented, only those candidate hits with Levenshtein distance equal to 0 are kept and assigned the maximum score (1). The OOV term detections found using this phone-based STD system are directly merged with the INV detections obtained using the word-based STD system.

3.10 System comparison

The systems submitted to the evaluation convey both similar and different properties that make them all interesting from a system comparison perspective. All the QbE STD systems employed DTW or DTW variants for the query search, for which the cost function is in general, the cosine distance. In addition, almost all the QbE STD systems employed fusion to output the final list of query detections. Regarding the feature extraction, the systems are based, in general, on posteriorgram-derived features for the query/utterance representation. However, there are specific differences that make each system distinct: The systems submitted by the ELiRF-UPV group (D-ELiRF-UPV-Post+DTW and E-ELiRF-UPV-Post+DTWNorm) differ in the cost matrix used within the S-DTW search. The systems submitted by the SPL-IT-UC group (F-SPL-IT-UC-4-phnrec+DTW fusion, G-SPL-IT-UC-3-phnrec+DTW fusion, and H-SPL-IT-UC-2-LIphnrec+DTW fusion) differ in the number of subsystems that are used for the fusion. The systems submitted by the L2F group (B-L2F-4-pllr fea+DTW fusion and C-L2F-4-likel fea+DTW fusion) show the most significant differences, both in the feature extractor, the DTW search, and the systems that are fused. Table 8 highlights the main differences and consistencies corresponding to the feature extraction, the cost functions, the search algorithm, and the fusion of each QbE STD system.

Table 8 Summary of the QbE STD systems

System ID	Feature extraction	Search	Cost function	Fusion
A-GTM-UVigo-3-fea+DTW fusion	Phoneme post.+unit sel. Acoustic feat.+feat. sel. Gaussian post.	S-DTW	Pearson corr. coef.	3 systems
B-L2F-4-pllr fea+DTW fusion	PLLR-phoneme-post.	DTW-thres1	Cosine distance	4 systems
C-L2F-4-likel fea+DTW fusion	LL-phoneme-post.	DTW-thres2	Cosine distance	4 systems (*)
D-ELiRF-UPV-Post+DTW	Phoneme post.	S-DTW	Cosine distance	–
E-ELiRF-UPV-Post+DTWNorm	Phoneme post.	S-DTW+Norm.	Cosine distance	–
F-SPL-IT-UC-4-phnrec+DTW fusion	Phoneme post.	DTW	Cosine distance	4 systems
G-SPL-IT-UC-3-phnrec+DTW fusion	Phoneme post.	DTW	Cosine distance	3 systems
H-SPL-IT-UC-2-LIphnrec+DTW fusion	Phoneme post.	DTW	Cosine distance	2 systems

post. posteriorgram, *sel.* selection, *feat.* features, *S-DTW* subsequence dynamic time warping, *corr. coef.* correlation coefficient, *PLLR* phone log-likelihood ratio, *DTW* dynamic time warping, *thres1* threshold 1, *thres2* threshold 2, *LL* log-likelihood, *Norm* normalization. (*) These 4 systems are different from those fused in the B-L2F-4-pllr fea+DTW fusion system

4 Results and discussion

The system results are presented in Table 9 for the development data, and Tables 10 and 11 show the performance for the MAVIR and the EPIC test data, respectively. The most important findings in the results are presented in Table 12.

4.1 Development data results

- The best performance for the QbE STD task was obtained by the C-L2F-4-likel fea+DTW fusion system. This system explicitly models the target language, i.e., Spanish, using a specific phoneme recognizer and is based on the fusion of different phoneme recognizers, since these improve the system performance. Paired t tests show that this best performance was statistically significant when compared with the B-L2F-4-pllr fea+DTW ($p < 0.02$), D-ELiRF-UPV-Post+DTW ($p < 0.01$), E-ELiRF-UPV-Post+DTWNorm ($p < 0.01$), and H-SPL-IT-UC-2-LIphnrec+DTW fusion ($p < 0.01$) systems.
- The worst performance was exhibited by the D-ELiRF-UPV-Post+DTW and E-ELiRF-UPV-Post+DTWnorm systems, which did not employ any fusion strategy.
- The performance obtained by the H-SPL-IT-UC-2-LIphnrec+DTW fusion system also confirmed significant performance degradation when the target language information was not used in the system. However, the A-GTM-UVigo-3-fea+DTW fusion system was an exception; although this did not employ the target language information, it still obtained a reasonable performance. This effect is possibly due to the use of a robust feature extractor, which involves the feature selection and the phoneme unit selection.
- As expected, the I-Text-based STD system, which employed the correct transcription of the query as input and the target language information, significantly outperformed all the QbE STD systems ($p < 0.01$). However, it must be noted that this I-Text-based STD system did not compete in the evaluation itself, because it did not follow the rules of the evaluation.

4.2 Test data results

4.2.1 MAVIR test data

- The system with the best performance for the QbE STD task does not match the system of the development data. On these test data, the best

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

System ID	MTWV	ATWV	p(FA)	p(Miss)	Fusion
A-GTM-UVigo-3-fea+DTW fusion	0.2800	0.2750	0.00002	0.699	YES
B-L2F-4-pllr fea+DTW fusion	0.2422	0.2247	0.00005	0.704	YES
C-L2F-4-likel fea+DTW fusion	0.3190	0.3099	0.00005	0.635	YES
D-ELiRF-UPV-Post+DTW	0.1991	0.1991	0.00000	0.801	NO
E-ELiRF-UPV-Post+DTWNorm	0.2057	0.2057	0.00000	0.794	NO
F-SPL-IT-UC-4-phnrec+DTW fusion	0.2954	0.2954	0.00008	0.621	YES
G-SPL-IT-UC-3-phnrec+DTW fusion	0.3001	0.3001	0.00002	0.683	YES
H-SPL-IT-UC-2-LIphnrec+DTW fusion	0.2009	0.2009	0.00005	0.749	YES
I-Text-based STD	0.6576	0.6559	0.00005	0.288	YES

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

System ID	MTWV	ATWV	p(FA)	p(Miss)	Fusion
A-GTM-UVigo-3-fea+DTW fusion	0.2739	0.2646	0.00008	0.651	YES
B-L2F-4-pllr fea+DTW fusion	0.2343	0.2287	0.00005	0.715	YES
C-L2F-4-likel fea+DTW fusion	0.2789	0.2542	0.00006	0.657	YES
D-ELiRF-UPV-Post+DTW	0.2003	0.1729	0.00002	0.779	NO
E-ELiRF-UPV-Post+DTWNorm	0.1958	0.1759	0.00003	0.776	NO
F-SPL-IT-UC-4-phnrec+DTW fusion	0.2674	0.2294	0.00005	0.685	YES
G-SPL-IT-UC-3-phnrec+DTW fusion	0.2682	0.2427	0.00005	0.679	YES
H-SPL-IT-UC-2-LIphnrec+DTW fusion	0.2137	0.1913	0.00005	0.736	YES
I-Text-based STD	0.6414	0.6260	0.00006	0.298	YES

performance was for the A-GTM-UVigo-3-fea+DTW fusion system. We consider this may be due to some over-adaptation of the selected phoneme recognizers for the query search and the fusion to the development data, which caused a worse generalization on unseen (test) data.

- The best performance of the A-GTM-UVigo-3-fea+DTW fusion system could be due to the *robust* feature extraction it employs. This system is language-independent and hence is suitable to build a language-independent STD system, which is a hot topic in the search-of-speech. The results obtained with this system suggest that a fusion strategy combined with a robust feature extractor, which integrates a varied set of features in individual search processes, can alleviate the gap between language-dependent and language-independent QbE STD systems in highly difficult domains such as spontaneous speech. This best performance was statistically significant for a paired t test compared with the D-ELiRF-UPV-Post+DTW ($p < 0.01$), E-ELiRF-UPV-Post+DTWNorm ($p < 0.01$) and H-SPL-IT-UC-2-LIphnrec+DTW fusion ($p < 0.01$) systems.
- The remainder of the findings observed in the development data can also be found in the test data: The worst systems did not employ the target language information nor fusion, and the I-Text-based STD system significantly outperformed the QbE STD systems ($p < 0.01$).

4.2.2 EPIC test data

- The best performance for the QbE STD task was for the language-dependent G-SPL-IT-UC-3-phnrec+DTW fusion system. We consider the discrepancy compared with the MAVIR database relies on the change of the acoustic domain. The parameter tuning and the ATWV threshold estimation could dramatically change the system performance ranking (as in the A-GTM-UVigo-3-fea+DTW fusion system) when different domain data are used for training/development and test. The best performance of the G-SPL-IT-UC-3-phnrec+DTW fusion system was statistically significant for a paired t test compared with the A-GTM-UVigo-3-fea+DTW fusion ($p < 0.01$), B-L2F-4-pllr fea+DTW ($p < 0.01$), D-ELiRF-UPV-Post+DTW ($p < 0.01$), and E-ELiRF-UPV-Post+DTWNorm ($p < 0.01$) systems, and weakly significant compared with the

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

System ID	MTWV	ATWV	p(FA)	p(Miss)	Fusion
A-GTM-UVigo-3-fea+DTW fusion	0.2496	−0.4356	0.00008	0.668	YES
B-L2F-4-pllr fea+DTW fusion	0.2243	0.2181	0.00009	0.690	YES
C-L2F-4-likel fea+DTW fusion	0.2973	0.2628	0.00012	0.587	YES
D-ELiRF-UPV-Post+DTW	0.1530	0.1103	0.00001	0.835	NO
E-ELiRF-UPV-Post+DTWNorm	0.1658	0.1232	0.00003	0.800	NO
F-SPL-IT-UC-4-phnrec+DTW fusion	0.3334	0.2641	0.00007	0.593	YES
G-SPL-IT-UC-3-phnrec+DTW fusion	0.3277	0.3011	0.00006	0.610	YES
H-SPL-IT-UC-2-LIphnrec+DTW fusion	0.2662	0.2500	0.00007	0.664	YES
I-Text-based STD	0.8617	0.8586	0.00004	0.097	YES

Table 12 Summary of the best QbE STD system results and the I-Text based-STD system of the ALBAYZIN QbE STD 2016 evaluation

Best system ID	ATWV	Data	Correct query transcription	Fusion	Lang-dep.
C-L2F-4-likel fea+DTW fusion	0.3099	Development data	NO	YES	YES
A-GTM-UVigo-3-fea+DTW fusion	0.2646	MAVIR test data	NO	YES	NO
G-SPL-IT-UC-3-phnrec+DTW fusion	0.3011	EPIC test data	NO	YES	YES
I-Text-based STD	0.6559	Development data	YES	YES	YES
I-Text-based STD	0.6260	MAVIR test data	YES	YES	YES
I-Text-based STD	0.8586	EPIC test data	YES	YES	YES

Lang-dep. language dependency

F-SPL-IT-UC-4-phnrec+DTW fusion and H-SPL-IT-UC-2-LIphnrec+DTW fusion ($p < 0.05$) systems. It must be noted that the significance levels decrease for the language-independent QbE STD systems due to the change of the acoustic domain.

- Although from an ASR perspective, the EPIC database is easier compared to the MAVIR database, not all the systems obtained a better performance than that on the MAVIR data due to the domain change. The fusion strategy played an important role in alleviating this issue, since the systems that do not employ any fusion strategy degrade their performance to a greater extent with respect to the MAVIR test data, whereas the systems that are based on fusion obtain similar or better performance than that obtained in the MAVIR test data.
- The A-GTM-UVigo-3-fea+DTW fusion system dramatically decreases the performance due to an issue in the estimation of the ATWV threshold.
- The results suggest that using the target language is not that beneficial when the acoustic domain of the development and the test data changes, since the performance of the language-independent QbE STD systems, i.e., H-SPL-IT-UC-2-LIphnrec+DTW fusion, is better than that of some language-dependent QbE STD systems, i.e., B-L2F-4-pllr fea+DTW fusion.

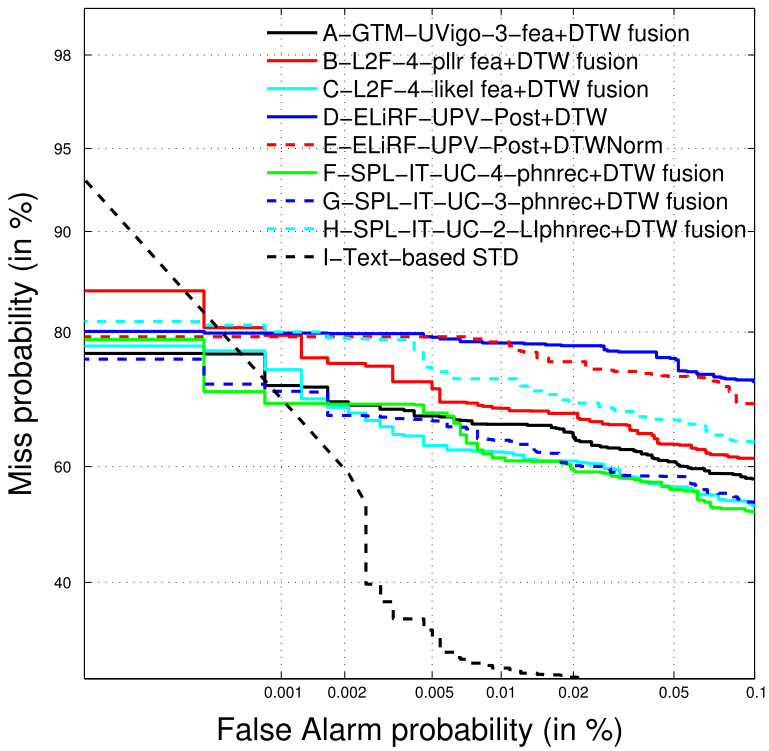

Fig. 6 DET curves of the QbE STD and the text-based STD systems on the development data

- The I-Text-based STD system, as in the other datasets, significantly outperformed the performance of the QbE STD systems ($p < 0.01$).

4.3 Development and test data DET curves

The DET curves are presented in Figs. 6, 7, and 8 for the development data, the MAVIR test data, and the EPIC test data, respectively. These show a similar pattern to that observed in the system ranking from the MTWV/ATWV results.

4.4 System performance analysis based on query length

An analysis of the system performance based on the length of the queries was carried out. The results are presented in Tables 13 and 14 for the MAVIR and the EPIC test data, respectively. For the MAVIR data, it can be observed that, in general, the long queries obtained the best performance. This is due to the fact that when the length of the query increases, there is less confusion between the query terms, because these typically differ to a great extent and hence a better performance is obtained. However, it can also be seen that the short queries outperformed the medium-length queries. This could be due to the fact that the short queries, which contain up to 7 phonemes, are not short enough to make the QbE STD performance worse compared to the medium-length queries, which contain between 8 and 10 phonemes. For the I-Text-based STD system, the medium-length queries obtained the best performance. These outperformed the short queries, because as described above, there is less acoustic confusion the longer the length of the query. In this I-Text-based STD system, the medium-length queries also performed better than the long queries, which may be related to the fact that the long queries have an OOV rate of 56%, whereas the medium-length queries have an OOV rate of 39%.

For the EPIC data, although the best performance also corresponded to the long queries, a different pattern of behaviour is observed: In general, the medium-length queries outperformed the short queries. This discrepancy with the MAVIR data may rely on the different conditions of each database such as the different number of queries, type of speech, and acoustic conditions. For the I-Text-based STD system, the long-length queries performed slightly better than the short- and medium-length queries, probably due to the lesser acoustic confusion. For this system, the medium-length queries performed slightly worse than the short-length queries. Although this may be surprising, it must be noted that some of the short-length queries can contain up to 7 phonemes, and so are not really very short.

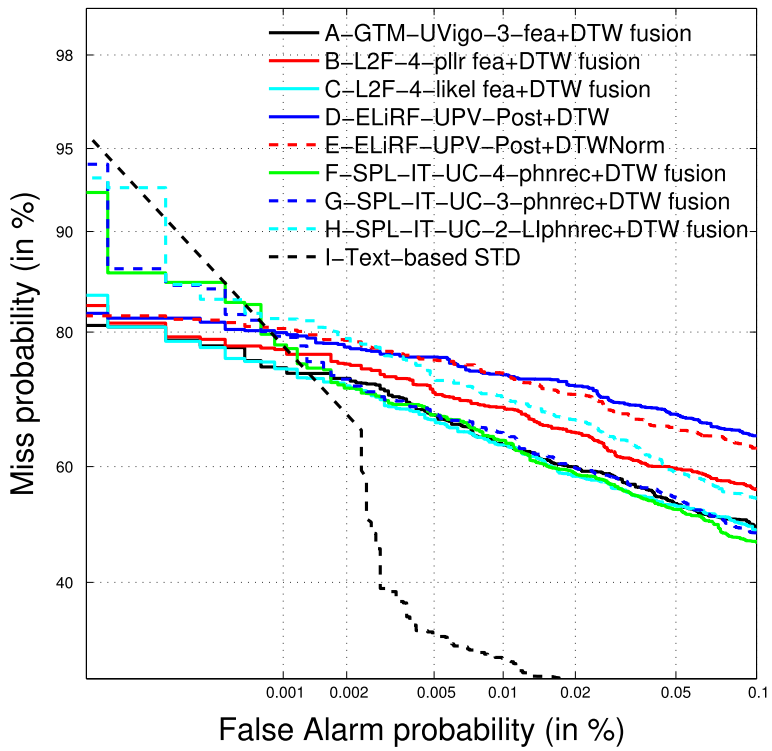

Fig. 7 DET curves of the QbE STD and the text-based STD systems on the MAVIR test data

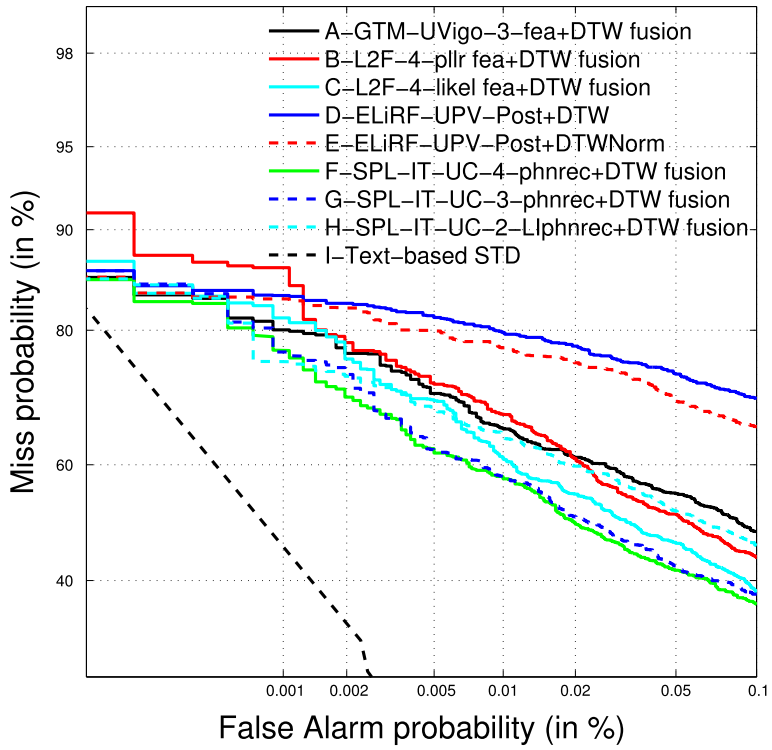

Fig. 8 DET curves of the QbE STD and the text-based STD systems on the EPIC test data

4.5 System performance analysis based on single-word/multi-word queries

An analysis of the system performance based on the single-word and the multi-word queries was carried out, and the results are presented in Table 15. The results show a degradation in performance from the multi-word to the single-word queries. The multi-word queries are typically longer than the single-word queries, and hence, better performance could be expected, as shown in the query length analysis. The only exception is the I-Text-based STD system, for which the ATWV performance was worse for the multi-word queries than for the single-word queries. However, it should be noted that the MTWV was much better for the multi-word queries. This indicates a problem in the threshold setting for multi-word queries.

4.6 System performance analysis based on in-language/out-of-language queries

An analysis of the system performance based on the in-language and the out-of-language queries was carried out and the results are presented in Table 16. These results

Table 13 System results of the ALBAYZIN QbE STD 2016 evaluation on the MAVIR test data based on the query length

System ID	Short		Medium		Long	
	MTWV	ATWV	MTWV	ATWV	MTWV	ATWV
A-GTM-UVigo-3-fea+DTW fusion	0.2976	0.2765	0.2353	0.2224	0.3388	0.3256
B-L2F-4-pllr fea+DTW fusion	0.2498	0.2346	0.1958	0.1850	0.3197	0.3009
C-L2F-4-likel fea+DTW fusion	0.2820	0.2529	0.2533	0.2326	0.3499	0.2957
D-ELiRF-UPV-Post+DTW	0.2219	0.2134	0.1889	0.1452	0.2319	0.1672
E-ELiRF-UPV-Post+DTWNorm	0.2219	0.2169	0.1804	0.1464	0.2455	0.1729
F-SPL-IT-UC-4-phnrec+DTW fusion	0.3011	0.2544	0.2291	0.1710	0.3318	0.3020
G-SPL-IT-UC-3-phnrec+DTW fusion	0.3013	0.2480	0.2209	0.1941	0.3332	0.3244
H-SPL-IT-UC-2-LIphnrec+DTW fusion	0.2296	0.2075	0.2078	0.1815	0.2143	0.1865
I-Text-based STD	0.5718	0.5472	0.7044	0.6959	0.6439	0.6077

Short short-length queries (up to 7 phonemes), *Medium* medium-length queries (between 8 and 10 phonemes), *Long* long-length queries (more than 10 phonemes)

Table 14 System results of the ALBAYZIN QbE STD 2016 evaluation on the EPIC test data based on the query length

System ID	Short		Medium		Long	
	MTWV	ATWV	MTWV	ATWV	MTWV	ATWV
A-GTM-UVigo-3-fea+DTW fusion	0.2228	−0.4374	0.2781	−0.3902	0.2312	−0.5509
B-L2F-4-pllr fea+DTW fusion	0.1825	0.1737	0.2355	0.2226	0.2967	0.2800
C-L2F-4-likel fea+DTW fusion	0.2943	0.2743	0.3044	0.2666	0.3265	0.2340
D-ELiRF-UPV-Post+DTW	0.1300	0.0870	0.1680	0.1139	0.1821	0.1399
E-ELiRF-UPV-Post+DTWNorm	0.1408	0.1008	0.1779	0.1268	0.2185	0.1510
F-SPL-IT-UC-4-phnrec+DTW fusion	0.2945	0.2321	0.3421	0.2761	0.4012	0.2862
G-SPL-IT-UC-3-phnrec+DTW fusion	0.2860	0.2561	0.3423	0.3191	0.4082	0.3289
H-SPL-IT-UC-2-LIphnrec+DTW fusion	0.2526	0.2253	0.2638	0.2332	0.3356	0.3352
I-Text-based STD	0.8712	0.8626	0.8501	0.8474	0.8847	0.8814

Short, Medium, and Long denote the same as in Table 13

show a degradation in performance from the out-of-language to the in-language queries. This is the reverse of what should be expected in the case of a language-dependent setup. However, since all the QbE STD systems rely on the fusion of search systems that employ different languages, the OOL issue becomes almost irrelevant. The OOL queries can obtain a better performance than the INL queries in a QbE STD system in the case where the OOL query language is employed to build the system. In this case, the English language was chosen for the OOL queries, and all the QbE STD systems (except the B-L2F-4-pllr fea+DTW fusion system) used English in the feature extraction module. Regarding the B-L2F-4-pllr fea+DTW fusion system, the fusion strategy still performs better on the OOL queries because four different languages are fused.

On the other hand, performance degradation is observed from the INL to the OOL queries in the I-Text-based STD system. In this case, the system is language-dependent because only the Spanish language was used to build the system, and hence a worse performance was obtained for the OOL queries because they did not match the target language. However, for the INL queries, where the pronunciation matches the target language, and for which enough data are typically used to train both the AMs and LMs, the system performance improved when compared to that of the QbE STD systems.

4.7 Comparison with the ALBAYZIN QbE STD 2014 evaluation

In order to measure the progress of the QbE STD technology in Spanish, a comparison of the best results obtained in the common set of queries of the ALBAYZIN QbE STD evaluations held in 2014 and 2016 was carried out. The best performance obtained in the 2014 and 2016 evaluations in the common set of queries was ATWV = 0.2881 and ATWV = 0.2541, respectively, which showed some performance degradation. It must be noted that the system submitted to the evaluation held in 2014 fuses the results of the text-based STD and the template matching-based approaches, which resulted in a better performance. On the contrary, the best system presented in the 2016

Table 15 System results of the ALBAYZIN QbE STD 2016 evaluation on the MAVIR test data for the single-word ('Single') and the multi-word ('Multi') queries

System ID	Single				Multi			
	MTWV	ATWV	p(FA)	p(Miss)	MTWV	ATWV	p(FA)	p(Miss)
A-GTM-UVigo-3-fea+DTW fusion	0.2467	0.2355	0.00008	0.675	0.7500	0.7500	0.00000	0.250
B-L2F-4-pllr fea+DTW fusion	0.2061	0.2001	0.00005	0.743	0.7042	0.7042	0.00005	0.250
C-L2F-4-likel fea+DTW fusion	0.2562	0.2355	0.00006	0.682	0.7500	0.5667	0.00000	0.250
D-ELiRF-UPV-Post+DTW	0.1737	0.1483	0.00002	0.806	0.6667	0.5833	0.00000	0.333
E-ELiRF-UPV-Post+DTWNorm	0.1689	0.1515	0.00003	0.802	0.6667	0.5833	0.00000	0.333
F-SPL-IT-UC-4-phnrec+DTW fusion	0.2356	0.1951	0.00008	0.680	0.8479	0.8021	0.00007	0.083
G-SPL-IT-UC-3-phnrec+DTW fusion	0.2371	0.2050	0.00005	0.709	0.8708	0.8708	0.00005	0.083
H-SPL-IT-UC-2-LIphnrec+DTW fusion	0.1907	0.1682	0.00005	0.760	0.6667	0.5750	0.00000	0.333
I-Text-based STD	0.6498	0.6385	0.00006	0.285	0.9167	0.4167	0.00000	0.083

Table 16 System results of the ALBAYZIN QbE STD 2016 evaluation on the MAVIR test data for the in-language (INL) and the out-of-language queries (OOL)

System ID	INL				OOL			
	MTWV	ATWV	p(FA)	p(Miss)	MTWV	ATWV	p(FA)	p(Miss)
A-GTM-UVigo-3-fea+DTW fusion	0.2627	0.2538	0.00005	0.687	0.5504	0.4168	0.00020	0.253
B-L2F-4-pllr fea+DTW fusion	0.2287	0.2213	0.00005	0.723	0.3714	0.3329	0.00000	0.629
C-L2F-4-likel fea+DTW fusion	0.2662	0.2365	0.00005	0.687	0.5275	0.5043	0.00008	0.394
D-ELiRF-UPV-Post+DTW	0.1935	0.1599	0.00002	0.788	0.3561	0.3561	0.00000	0.644
E-ELiRF-UPV-Post+DTWNorm	0.1872	0.1632	0.00003	0.785	0.3561	0.3561	0.00000	0.644
F-SPL-IT-UC-4-phnrec+DTW fusion	0.2538	0.2091	0.00005	0.696	0.5561	0.5168	0.00016	0.287
G-SPL-IT-UC-3-phnrec+DTW fusion	0.2564	0.2332	0.00005	0.692	0.4972	0.3768	0.00022	0.287
H-SPL-IT-UC-2-LIphnrec+DTW fusion	0.2075	0.1866	0.00005	0.740	0.3433	0.2571	0.00035	0.303
I-Text-based STD	0.6695	0.6576	0.00004	0.286	0.2500	0.1786	0.00000	0.750

evaluation was language-independent and included only template matching approaches. The data employed for the training and development varied from one evaluation to another. In the 2016 evaluation, there were less training data belonging to the MAVIR domain and the participants could not use the same data for training and development which could have influenced the system performance gap. Nevertheless, the best result obtained in the 2016 evaluation is still remarkable, as it was obtained by a language-independent QbE STD system and did not employ text-based STD technology.

4.8 Towards a language-independent STD system

The feasibility of language-independent STD systems can be examined from the systems submitted to the ALBAYZIN QbE STD 2016 evaluation. By comparing the best language-independent QbE STD system (A-GTM-UVigo-3-fea+DTW fusion for the MAVIR data and H-SPL-IT-UC-2-LIphnrec+DTW fusion for the EPIC data) with the I-Text-based STD system, we can claim that building a language-independent STD system with a performance similar to that of a language-dependent STD system remains a challenge. This means that researchers still need to focus more on the QbE STD technology to approximate the language-independent to the language-dependent STD systems.

5 Conclusions

This paper presents the systems submitted to the ALBAYZIN QbE STD 2016 evaluation together with a text-based STD system for comparison purposes. Four different research groups took part in the evaluation, and eight different systems were submitted in total. All the systems submitted allowed INV and OOV query detection, because they were based on template matching techniques. With regard to the most novel and interesting technical contributions, the feature extraction employed in the A-GTM-UVigo-3-fea+DTW fusion system is worth mentioning. It uses three feature extraction methods that integrate different information sources and two different feature selection approaches. The B-L2F-4-pllr fea+DTW fusion system also presents a valuable feature extraction approach by computing phone log-likelihood ratios from two different phoneme recognizers. The candidate hit selection proposed in the F-SPL-IT-UC-4-phnrec+DTW fusion system is also worth mentioning.

The results showed that system fusion plays an important role in the QbE STD systems and that the language-independence issue can be partially compensated by using a robust feature extractor. Regarding the domain comparison, we showed that for an easy domain such as that of the EPIC data, with an easy query list, i.e., INV, INL, and single-word queries, even though the training and the development data belonged to a different domain, the performance was better (ATWV $=$ 0.3011) than that for MAVIR data (ATWV $=$ 0.2646), which presented a more difficult speech and query list and the same type of training and development data. The out-of-language query detection can obtain similar or even better performance than the in-language query detection when the language of those OOL queries is used to construct the system or the system fuses several language-dependent QbE STD systems. In addition, we also showed that multi-word query detection is *easier* than single-word query detection because the multi-word queries are generally longer than the single-word queries and that the long-length queries typically perform better.

A comparison of the results of the language-independent QbE STD system with those of the language-dependent text-based STD system presented in this paper shows that it is clear that there is still ample room for improvement to approximate the performance of a language-independent QbE STD system to that of a language-dependent text-based STD system. This

encourages the organizers to maintain this evaluation in the next ALBAYZIN evaluation campaign for which two different domains (including MAVIR data), and a cross-search, i.e., searching the development queries in the test speech data and searching the test queries in the development speech data, will be considered as a measure of the generalization capability of the systems to unseen data.

Endnotes

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

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

[3] http://catalog.elra.info/product_info.php?products_id=1145 (European Parliament Interpretation Corpus)

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

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

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

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

[8] https://github.com/cmusphinx/g2p-seq2seq

[9] https://sourceforge.net/projects/cmusphinx/files/Acoustic%20and%20Language%20Models/Spanish/

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

Acknowledgements
This work was partially supported by Fundação para a Ciência e Tecnologia (FCT) under the projects UID/EEA/50008/2013 (pluriannual funding in the scope of the LETSREAD project) and UID/CEC/50021/2013, and Grant SFRH/BD/97187/2013. Jorge Proença is supported by the SFRH/BD/97204/2013 FCT Grant. This work was also supported by the Galician Government ('Centro singular de investigación de Galicia' accreditation 2016-2019 ED431G/01 and the research contract GRC2014/024 (Modalidade: Grupos de Referencia Competitiva 2014)), the European Regional Development Fund (ERDF), the projects "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) and the TIN2015-64282-R funded by Ministerio de Economía y Competitividad in Spain, the Spanish Government through the project "TraceThem" (TEC2015-65345-P), and AtlantTIC ED431G/04.

Authors' contributions
JT and DTT designed, prepared, and were the organizers of the ALBAYZIN Query-by-example Spoken Term Detection 2016 evaluation. They also carried out the detailed analysis of the evaluation results presented in this paper. PL-O and LD-F built the A-GTM-UVigo-3-fea+DTW fusion system. JP and FP built the F-SPL-IT-UC-4-phnrec+DTW fusion, the G-SPL-IT-UC-3-phnrec+DTW fusion, and the H-SPL-IT-UC-2-LIphnrec+DTW fusion systems. FGG and ES built the D-ELiRF-UPV-Post+DTW and the E-ELiRF-UPV-Post+DTWNorm systems. AP and AA built the B-L2F-4-pllr fea+DTW fusion and the C-L2F-4-likel fea+DTW fusion systems. All the authors also contribute to the discussion of the system results. The main contributions of this paper are as follows: The systems submitted to the third Query-by-example Spoken Term Detection evaluation for Spanish language are presented. Increasing complexity in the query list from the previous Query-by-example Spoken Term Detection evaluations. An analysis of the system results from various query characteristics is presented. An analysis of the system results from two different domains is presented. All authors read and approved the final manuscript.

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

Publisher's Note

Author details
[1]Escuela Politécnica Superior. Universidad San Pablo-CEU, CEU Universities, Campus de Montepríncipe, Madrid, Spain. [2]AuDIaS, Universidad Autónoma de Madrid, Av. Francisco Tomás y Valiente, 11. Escuela Politécnica Superior, Madrid, Spain. [3]Universidade da Coruña, IRLab, CITIC, Campus de Elviña s/n, A Coruña, Spain. [4]Multimedia Technologies Group (GTM), AtlantTIC Research Center, E. E. Telecomunicación, Campus Universitario de Vigo, s/n, Vigo, Spain. [5]Instituto de Telecomunicações, Department of Electrical and Computer Engineering, University of Coimbra, Paço das Escolas, Coimbra, Portugal. [6]ELiRF - Departament de Sistemes Informàtics i Computació, Universitat Politècnica de València, Camino de Vera, s/n, Valencia, Spain. [7]L^2F - Spoken Language Systems Lab, INESC-ID, IST - Instituto Superior Técnico, University of Lisbon, Rua Alves Redol, 9, Lisbon, Portugal.

Received: 9 October 2017 Accepted: 25 March 2018
Published online: 13 April 2018

References

1. Jarina, R, Kuba, M, Gubka, R, Chmulik, M, Paralic, M (2013). UNIZA system for the spoken web search task at MediaEval 2013, In *Proc. of MediaEval* (pp. 791–792). New Haven: Ruzica Piskac.
2. Ali, A, & Clements, MA (2013). Spoken web search using and ergodic hidden Markov model of speech, In *Proc. of MediaEval* (pp. 861–862). New Haven: Ruzica Piskac.
3. Buzo, A, Cucu, H, Burileanu, C (2014). SpeeD@MediaEval 2014: Spoken term detection with robust multilingual phone recognition, In *Proc. of MediaEval* (pp. 721–722). New Haven: Ruzica Piskac.
4. Caranica, A, Buzo, A, Cucu, H, Burileanu, C (2015). SpeeD@MediaEval 2015: Multilingual phone recognition approach to Query By Example STD, In *Proc. of MediaEval* (pp. 781–783). New Haven: Ruzica Piskac.
5. Kesiraju, S, Mantena, G, Prahallad, K (2014). IIIT-H system for MediaEval 2014 QUESST, In *Proc. of MediaEval* (pp. 761–762). New Haven: Ruzica Piskac.
6. Ma, M, & Rosenberg, A (2015). CUNY systems for the Query-by-Example search on speech task at MediaEval 2015, In *Proc. of MediaEval* (pp. 831–833). New Haven: Ruzica Piskac.
7. Takahashi, J, Hashimoto, T, Konno, R, Sugawara, S, Ouchi, K, Oshima, S, Akyu, T, Itoh, Y (2014). An IWAPU STD system for OOV query terms and spoken queries, In *Proc. of NTCIR-11* (pp. 384–389). Tokyo: National Institute of Informatics.
8. Makino, M, & Kai, A (2014). Combining subword and state-level dissimilarity measures for improved spoken term detection in NTCIR-11 SpokenQuery & Doc task, In *Proc. of NTCIR-11* (pp. 413–418). Tokyo: National Institute of Informatics.
9. Konno, R, Ouchi, K, Obara, M, Shimizu, Y, Chiba, T, Hirota, T, Itoh, Y (2016). An STD system using multiple STD results and multiple rescoring method for NTCIR-12 SpokenQuery & Doc task, In *Proc. of NTCIR-12* (pp. 200–204). Tokyo: National Institute of Informatics.
10. Sakamoto, N, Yamamoto, K, Nakagawa, S (2015). Combination of syllable based N-gram search and word search for spoken term detection through spoken queries and IV/OOV classification, In *Proc. of ASRU* (pp. 200–206). New York: IEEE.
11. Hou, J, Pham, VT, Leung, C-C, Wang, L, 2, HX, Lv, H, Xie, L, Fu, Z, Ni, C, Xiao, X, Chen, H, Zhang, S, Sun, S, Yuan, Y, Li, P, Nwe, TL, Sivadas, S, Ma, B, Chng, ES, Li, H (2015). The NNI Query-by-Example system for MediaEval 2015, In *Proc. of MediaEval* (pp. 141–143). New Haven: Ruzica Piskac.
12. Vavrek, J, Viszlay, P, Lojka, M, Pleva, M, Juhar, J, Rusko, M (2015). TUKE at MediaEval 2015 QUESST, In *Proc. of MediaEval* (pp. 451–453). New Haven: Ruzica Piskac.
13. Mantena, G, Achanta, S, Prahallad, K (2014). Query-by-example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *22*(5), 946–955.
14. Anguera, X, & Ferrarons, M (2013). Memory efficient subsequence DTW for query-by-example spoken term detection, In *Proc. of ICME* (pp. 1–6). New York: IEEE.

15. Tulsiani, H, & Rao, P (2015). The IIT-B Query-by-Example system for MediaEval 2015, In *Proc. of MediaEval* (pp. 341–343). New Haven: Ruzica Piskac.
16. Bouallegue, M, Senay, G, Morchid, M, Matrouf, D, Linares, G, Dufour, R (2013). LIA@MediaEval 2013 spoken web search task: An I-Vector based approach, In *Proc. of MediaEval* (pp. 771–772). New Haven: Ruzica Piskac.
17. Rodriguez-Fuentes, LJ, Varona, A, Penagarikano, M, Bordel, G, Diez, M (2013). GTTS systems for the SWS task at MediaEval 2013, In *Proc. of MediaEval* (pp. 831–832). New Haven: Ruzica Piskac.
18. Wang, H, Lee, T, Leung, C-C, Ma, B, Li, H (2013). Using parallel tokenizers with DTW matrix combination for low-resource spoken term detection, In *Proc. of ICASSP* (pp. 8545–8549). New York: IEEE.
19. Wang, H, & Lee, T (2013). The CUHK spoken web search system for MediaEval 2013, In *Proc. of MediaEval* (pp. 681–682). New Haven: Ruzica Piskac.
20. Proenca, J, Veiga, A, Perdigão, F (2014). The SPL-IT query by example search on speech system for MediaEval 2014, In *Proc. of MediaEval* (pp. 741–742). New Haven: Ruzica Piskac.
21. Proenca, J, Veiga, A, Perdigao, F (2015). Query by example search with segmented dynamic time warping for non-exact spoken queries, In *Proc. of EUSIPCO* (pp. 1691–1695). Berlin: Springer.
22. Proenca, J, Castela, L, Perdigao, F (2015). The SPL-IT-UC Query by Example search on speech system for MediaEval 2015, In *Proc. of MediaEval* (pp. 471–473). New Haven: Ruzica Piskac.
23. Proenca, J, & Perdigao, F (2016). Segmented dynamic time warping for spoken Query-by-Example search, In *Proc. of Interspeech* (pp. 750–754). Baixas: ISCA.
24. Lopez-Otero, P, Docio-Fernandez, L, Garcia-Mateo, C (2015). GTM-UVigo systems for the Query-by-Example search on speech task at MediaEval 2015, In *Proc. of MediaEval* (pp. 521–523). New Haven: Ruzica Piskac.
25. Lopez-Otero, P, Docio-Fernandez, L, Garcia-Mateo, C (2015). Phonetic unit selection for cross-lingual Query-by-Example spoken term detection, In *Proc. of ASRU* (pp. 223–229). New York: IEEE.
26. Saxena, A, & Yegnanarayana, B (2015). Distinctive feature based representation of speech for Query-by-Example spoken term detection, In *Proc. of Interspeech* (pp. 3680–3684). Baixas: ISCA.
27. Skacel, M, & Szöke, I (2015). BUT QUESST 2015 system description, In *Proc. of MediaEval* (pp. 721–723). New Haven: Ruzica Piskac.
28. Chen, H, Leung, C-C, Xie, L, Ma, B, Li, H (2016). Unsupervised bottleneck features for low-resource Query-by-Example spoken term detection, In *Proc. of Interspeech* (pp. 923–927). Baixas: ISCA.
29. Yuan, Y, Leung, C-C, Xie, L, Chen, H, Ma, B, Li, H (2017). Pairwise learning using multi-lingual bottleneck features for low-resource Query-by-Example spoken term detection, In *Proc. of ICASSP* (pp. 5645–5649). New York: IEEE.
30. Torbati, AHHN, & Picone, J (2016). A nonparametric Bayesian approach for spoken term detection by example query, In *Proc. of Interspeech* (pp. 928–932). Baixas: ISCA.
31. Popli, A, & Kumar, A (2015). Query-by-example spoken term detection using low dimensional posteriorgrams motivated by articulatory classes, In *Proc. of MMSP* (pp. 1–6). New York: IEEE.
32. Yang, P, Leung, C-C, Xie, L, Ma, B, Li, H (2014). Intrinsic spectral analysis based on temporal context features for query-by-example spoken term detection, In *Proc. of Interspeech* (pp. 1722–1726). Baixas: ISCA.
33. George, B, Saxena, A, Mantena, G, Prahallad, K, Yegnanarayana, B (2014). Unsupervised query-by-example spoken term detection using bag of acoustic words and non-segmental dynamic time warping, In *Proc. of Interspeech* (pp. 1742–1746). Baixas: ISCA.
34. Hazen, TJ, Shen, W, White, CM (2009). Query-by-example spoken term detection using phonetic posteriorgram templates, In *Proc. of ASRU* (pp. 421–426). New York: IEEE.
35. Abad, A, Astudillo, RF, Trancoso, I (2013). The L2F spoken web search system for mediaeval 2013, In *Proc. of MediaEval* (pp. 851–852). New Haven: Ruzica Piskac.
36. Szöke, I, Skácel, M, Burget, L (2014). BUT QUESST 2014 system description, In *Proc. of MediaEval* (pp. 621–622). New Haven: Ruzica Piskac.
37. Szöke, I, Burget, L, Grézl, F, Černocký, JH, Ondel, L (2014). Calibration and fusion of query-by-example systems - BUT SWS 2013, In *Proc. of ICASSP* (pp. 621–622). New York: IEEE.
38. Abad, A, Rodríguez-Fuentes, LJ, Penagarikano, M, Varona, A, Bordel, G (2013). On the calibration and fusion of heterogeneous spoken term detection systems, In *Proc. of Interspeech* (pp. 20–24). Baixas: ISCA.
39. Yang, P, Xu, H, Xiao, X, Xie, L, Leung, C-C, Chen, H, Yu, J, Lv, H, Wang, L, Leow, SJ, Ma, B, Chng, ES, Li, H (2014). The NNI query-by-example system for MediaEval 2014, In *Proc. of MediaEval* (pp. 691–692). New Haven: Ruzica Piskac.
40. Leung, C-C, Wang, L, Xu, H, Hou, J, Pham, VT, Lv, H, Xie, L, Xiao, X, Ni, C, Ma, B, Chng, ES, Li, H (2016). Toward high-performance language-independent Query-by-Example spoken term detection for MediaEval 2015: Post-evaluation analysis, In *Proc. of Interspeech* (pp. 3703–3707). Baixas: ISCA.
41. Xu, H, Hou, J, Xiao, X, Pham, VT, Leung, C-C, Wang, L, Do, VH, Lv, H, Xie, L, Ma, B, Chng, ES, Li, H (2016). Approximate search of audio queries by using DTW with phone time boundary and data augmentation, In *Proc. of ICASSP* (pp. 6030–6034). New York: IEEE.
42. Oishi, S, Matsuba, T, Makino, M, Kai, A (2016). Combining state-level and DNN-based acoustic matches for efficient spoken term detection in NTCIR-12 SpokenQuery&Doc-2 task, In *Proc. of NTCIR-12* (pp. 205–210). Tokyo: National Institute of Informatics.
43. Oishi, S, Matsuba, T, Makino, M, Kai, A (2016). Combining state-level spotting and posterior-based acoustic match for improved query-by-example spoken term detection, In *Proc. of Interspeech* (pp. 740–744). Baixas: ISCA.
44. Obara, M, Kojima, K, Tanaka, K, Lee, S-w, Itoh, Y (2016). Rescoring by combination of posteriorgram score and subword-matching score for use in Query-by-Example, In *Proc. of Interspeech* (pp. 1918–1922). Baixas: ISCA.
45. NIST. The Ninth Text REtrieval Conference (TREC 9). http://trec.nist.gov. Accessed Feb 2018.
46. Anguera, X, Rodriguez-Fuentes, LJ, Szöke, I, Buzo, A, Metze, F (2014). Query by Example Search on Speech at Mediaeval 2014, In *Proc. of MediaEval* (pp. 351–352). New Haven: Ruzica Piskac.
47. Joho, H, & Kishida, K (2014). Overview of the NTCIR-11 SpokenQuery&Doc Task, In *Proc. of NTCIR-11* (pp. 1–7). Tokyo: National Institute of Informatics.
48. NIST. Draft KWS16 Keyword Search Evaluation Plan. https://www.nist.gov/sites/default/files/documents/itl/iad/mig/KWS16-evalplan-v04.pdf. Accessed Feb 2018.
49. Anguera, X, Metze, F, Buzo, A, Szöke, I, Rodriguez-Fuentes, LJ (2013). The spoken web search task, In *Proc. of MediaEval* (pp. 921–922). New Haven: Ruzica Piskac.
50. Anguera, X, Rodriguez-Fuentes, LJ, Szöke, I, Buzo, A, Metze, F (2014). Query by example search on speech at Mediaeval 2014, In *Proc. of MediaEval* (pp. 351–352). New Haven: Ruzica Piskac.
51. Taras, B, & Nadeu, C (2011). Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion. *EURASIP Journal on Audio, Speech, and Music Processing*, *2011*(1), 1–10.
52. Zelenák, M, Schulz, H, Hernando, J (2012). Speaker diarization of broadcast news in Albayzin 2010 evaluation campaign. *EURASIP Journal on Audio, Speech, and Music Processing*, *2012*(19), 1–9.
53. Rodríguez-Fuentes, LJ, Penagarikano, M, Varona, A, Díez, M, Bordel, G (2011). The Albayzin 2010 Language Recognition Evaluation, In *Proc. of Interspeech* (pp. 1529–1532). Baixas: ISCA.
54. Tejedor, J, Toledano, DT, Lopez-Otero, P, Docio-Fernandez, L, Garcia-Mateo, C, Cardenal, A, Echeverry-Correa, JD, Coucheiro-Limeres, A, Olcoz, J, Miguel, A (2015). Spoken term detection ALBAYZIN 2014 evaluation: overview, systems, results, and discussion. *EURASIP, Journal on Audio, Speech and Music Processing*, *2015*(21), 1–27.
55. Tejedor, J, Toledano, DT, Anguera, X, Varona, A, Hurtado, LF, Miguel, A, Colás, J (2013). Query-by-example spoken term detection ALBAYZIN 2012 evaluation: overview, systems, results, and discussion. *EURASIP, Journal on Audio, Speech, and Music Processing*, *2013*(23), 1–17.
56. Tejedor, J, Toledano, DT, Lopez-Otero, P, Docio-Fernandez, L, Garcia-Mateo, C (2016). Comparison of ALBAYZIN query-by-example spoken term detection 2012 and 2014 evaluations. *EURASIP, Journal on Audio, Speech and Music Processing*, *2016*(1), 1–19.
57. Méndez, F, Docío, L, Arza, M, Campillo, F (2010). The Albayzin 2010 text-to-speech evaluation, In *Proc. of FALA* (pp. 317–340). Vigo: UniversidadeVigo.

58. Billa, J, Ma, KW, McDonough, JW, Zavaliagkos, G, Miller, DR, Ross, KN, El-Jaroudi, A (1997). Multilingual speech recognition: the 1996 Byblos Callhome system, In *Proc. of Eurospeech* (pp. 363–366). Baixas: ISCA.
59. Killer, M, Stuker, S, Schultz, T (2003). Grapheme based speech recognition, In *Proc. of Eurospeech* (pp. 3141–3144). Baixas: ISCA.
60. Burget, L, Schwarz, P, Agarwal, M, Akyazi, P, Feng, K, Ghoshal, A, Glembek, O, Goel, N, Karafiat, M, Povey, D, Rastrow, A, Rose, RC, Thomas, S (2010). Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models, In *Proc. of ICASSP* (pp. 4334–4337). New York: IEEE.
61. Cuayahuitl, H, & Serridge, B (2002). Out-of-vocabulary word modeling and rejection for Spanish keyword spotting systems, In *Proc. of MICAI* (pp. 156–165). Berlin: Springer.
62. Tejedor, J (2009). Contributions to keyword spotting and spoken term detection for information retrieval in audio mining. PhD thesis, Universidad Autónoma de Madrid, Madrid, Spain.
63. Tejedor, J, Toledano, DT, Wang, D, King, S, Colás, J (2014). Feature analysis for discriminative confidence estimation in spoken term detection. *Computer Speech and Language*, *28*(5), 1083–1114.
64. Li, J, Wang, X, Xu, B (2014). An empirical study of multilingual and low-resource spoken term detection using deep neural networks, In *Proc. of Interspeech* (pp. 1747–1751). Baixas: ISCA.
65. NIST. The Spoken Term Detection (STD) 2006 evaluation plan. http://berlin.csie.ntnu.edu.tw/Courses/Special%20Topics%20in%20Spoken%20Language%20Processing/Lectures2008/SLP2008S-Lecture12-Spoken%20Term%20Detection.pdf. Accessed Feb 2018.
66. Fiscus, JG, Ajot, J, Garofolo, JS, Doddingtion, G (2007). Results of the 2006 spoken term detection evaluation, In *Proc. of SSCS* (pp. 45–50). New York: ACM.
67. Martin, A, Doddington, G, Kamm, T, Ordowski, M, Przybocki, M (1997). The DET curve in assessment of detection task performance, In *Proc. of Eurospeech* (pp. 1895–1898). Baixas: ISCA.
68. NIST. Evaluation Toolkit (STDEval) software. https://www.nist.gov/itl/iad/mig/tools. Accessed Feb 2018.
69. Union, IT. 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 Feb 2018.
70. Rajput, N, & Metze, F (2011). Spoken web search, In *Proc. of MediaEval* (pp. 1–2). New Haven: Ruzica Piskac.
71. Metze, F, Barnard, E, Davel, M, van Heerden, C, Anguera, X, Gravier, G, Rajput, N (2012). The spoken web search task, In *Proc. of MediaEval* (pp. 41–42). New Haven: Ruzica Piskac.
72. Szöke, I, Rodriguez-Fuentes, LJ, Buzo, A, Anguera, X, Metze, F, Proenca, J, Lojka, M, Xiong, X (2015). Query by Example Search on Speech at Mediaeval 2015, In *Proc. of MediaEval* (pp. 81–82). New Haven: Ruzica Piskac.
73. Szöke, I, & Anguera, X (2016). Zero-cost speech recognition task at Mediaeval 2016, In *Proc. of MediaEval* (pp. 81–82). New Haven: Ruzica Piskac.
74. Akiba, T, Nishizaki, H, Nanjo, H, Jones, GJF (2014). Overview of the NTCIR-11 spokenquery&doc task, In *Proc. of NTCIR-11* (pp. 1–15). Tokyo: National Institute of Informatics.
75. Akiba, T, Nishizaki, H, Nanjo, H, Jones, GJF (2016). Overview of the NTCIR-12 spokenquery&doc-2, In *Proc. of NTCIR-12* (pp. 1–13). Tokyo: National Institute of Informatics.
76. Schwarz, P (2008). Phoneme recognition based on long temporal context. PhD thesis, FIT, BUT, Brno, Czech Republic.
77. Varona, A, Penagarikano, M, Rodríguez-Fuentes, LJ, Bordel, G (2011). On the use of lattices of time-synchronous cross-decoder phone co-occurrences in a SVM-phonotactic language recognition system, In *Proc. of Interspeech* (pp. 2901–2904). Baixas: ISCA.
78. Eyben, F, Wollmer, M, Schuller, B (2010). OpenSMILE—the munich versatile and fast open-source audio feature extractor, In *Proc. of ACM Multimedia (MM)* (pp. 1459–1462). New York: ACM.
79. Lopez-Otero, P, Docio-Fernandez, L, Garcia-Mateo, C (2016). Finding relevant features for zero-resource query-by-example search on speech. *Speech Communication*, *84*(1), 24–35.
80. Zhang, Y, & Glass, JR (2009). Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams, In *Proc. of ASRU* (pp. 398–403). New York: IEEE.
81. Povey, D, Ghoshal, A, Boulianne, G, Burget, L, Glembek, O, Goel, N, Hannemann, M, Motlicek, P, Qian, Y, Schwarz, P, Silovsky, J, Stemmer, G, Vesely, K (2011). The KALDI speech recognition toolkit, In *Proc. of ASRU* (pp. 1–4). New York: IEEE.
82. Muller, M (2007). *Information retrieval for music and motion*. New York: Springer.
83. Szöke, I, Skacel, M, Burget, L (2014). BUT QUESST 2014 system description, In *Proc. of MediaEval* (pp. 621–622). New Haven: Ruzica Piskac.
84. Brümmer, N, & van Leeuwen, D (2006). On calibration of language recognition scores, In *Proc of the IEEE Odyssey: The speaker and language recognition workshop* (pp. 1–8). New York: IEEE.
85. Brümmer, N, & de Villiers, E. The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing. Technical report. https://sites.google.com/site/nikobrummer. Accessed Feb 2018.
86. Meinedo, H, & Neto, J (2005). A stream-based audio segmentation, classification and clustering pre-processing system for broadcast news using ANN models, In *Proc. of Interspeech* (pp. 237–240). Baixas: ISCA.
87. Morgan, N, & Bourlard, H (1995). An introduction to hybrid HMM/connectionist continuous speech recognition. *IEEE Signal Processing Magazine*, *12*(3), 25–42.
88. Meinedo, H, Abad, A, Pellegrini, T, Trancoso, I, Neto, J (2010). The L2F broadcast news speech recognition system, In *Proc. of FALA* (pp. 93–96). Vigo: UniversidadeVigo.
89. Abad, A, Luque, J, Trancoso, I (2011). Parallel transformation network features for speaker recognition, In *Proc. of ICASSP* (pp. 5300–5303). New York: IEEE.
90. Diez, M, Varona, A, Penagarikano, M, Rodriguez-Fuentes, LJ, Bordel, G (2012). On the use of phone log-likelihood ratios as features in spoken language recognition, In *Proc. of SLT* (pp. 274–279). New York: IEEE.
91. Diez, M, Varona, A, Penagarikano, M, Rodriguez-Fuentes, LJ, Bordel, G (2014). New insight into the use of phone log-likelihood ratios as features for language recognition, In *Proc. of Interspeech* (pp. 1841–1845). Baixas: ISCA.
92. Abad, A, Ribeiro, E, Kepler, F, Astudillo, R, Trancoso, I (2016). Exploiting phone log-likelihood ratio features for the detection of the native language of non-native English speakers, In *Proc. of Interspeech* (pp. 2413–2417). Baixas: ISCA.
93. Rodríguez-Fuentes, LJ, Varona, A, Peñagarikano, M, Bordel, G, Díez, M (2014). High-performance query-by-example spoken term detection on the SWS 2013 evaluation, In *Proc. of ICASSP* (pp. 7819–7823). New York: IEEE.
94. Proenca, J, & Perdigao, F (2016). Segmented dynamic time warping for spoken Query-by-Example search, In *Proc. of Interspeech* (pp. 750–754). Baixas: ISCA.
95. Vesely, K, Ghoshal, A, Burget, L, Povey, D (2013). Sequence-discriminative training of deep neural networks, In *Proc. of Interspeech* (pp. 2345–2349). Baixas: ISCA.
96. Ghahremani, P, BabaAli, B, Povey, D, Riedhammer, K, Trmal, J, Khudanpur, S (2014). A pitch extraction algorithm tuned for automatic speech recognition, In *Proc. of ICASSP* (pp. 2494–2498). New York: IEEE.
97. Povey, D, Hannemann, M, Boulianne, G, Burget, L, Ghoshal, A, Janda, M, Karafiat, M, Kombrink, S, Motlicek, P, Qian, Y, Riedhammer, K, Vesely, K, Vu, NT (2012). Generating exact lattices in the WFST framework, In *Proc. of ICASSP* (pp. 4213–4216). New York: IEEE.
98. Garcia-Mateo, C, Dieguez-Tirado, J, Docio-Fernandez, L, Cardenal-Lopez, A (2004). Transcrigal: A bilingual system for automatic indexing of broadcast news, In *Proc. of LREC* (pp. 2061–2064). Paris: ELRA.
99. Stolcke, A (2002). SRILM—an extensible language modeling toolkit, In *Proc. of Interspeech* (pp. 901–904). Baixas: ISCA.
100. Lopez-Otero, P, Docio-Fernandez, L, Garcia-Mateo, C (2016). GTM-UVigo systems for Albayzin 2016 search on speech evaluation, In *Proc. of Iberspeech* (pp. 65–74). Berlin: Springer.
101. Chen, G, Khudanpur, S, Povey, D, Trmal, J, Yarowsky, D, Yilmaz, O (2013). Quantifying the value of pronunciation lexicons for keyword search in low resource languages, In *Proc. of ICASSP* (pp. 8560–8564). New York: IEEE.
102. Pham, VT, Chen, NF, Sivadas, S, Xu, H, Chen, I-F, Ni, C, Chng, ES, Li, H (2014). System and keyword dependent fusion for spoken term detection, In *Proc. of SLT* (pp. 430–435). New York: IEEE.

103. Can, D, & Saraclar, M (2011). Lattice indexing for spoken term detection. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(8), 2338–2347.
104. Miller, DRH, Kleber, M, Kao, C-L, Kimball, O, Colthurst, T, Lowe, SA, Schwartz, RM, Gish, H (2007). Rapid and accurate spoken term detection, In *Proc. of Interspeech* (pp. 314–317). Baixas: ISCA.