Tejedor *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2015) 2015:21
DOI 10.1186/s13636-015-0063-8

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Spoken term detection ALBAYZIN 2014 evaluation: overview, systems, results, and discussion

Javier Tejedor[1*], Doroteo T. Toledano[2], Paula Lopez-Otero[3], Laura Docio-Fernandez[3], Carmen Garcia-Mateo[3], Antonio Cardenal[3], Julian David Echeverry-Correa[4], Alejandro Coucheiro-Limeres[4], Julia Olcoz[5] and Antonio Miguel[5]

Abstract

Spoken term detection (STD) aims at retrieving data from a speech repository given a textual representation of the search term. Nowadays, it is receiving much interest due to the large volume of multimedia information. STD differs from automatic speech recognition (ASR) in that ASR is interested in all the terms/words that appear in the speech data, whereas STD focuses on a selected list of search terms that must be detected within the speech data. This paper presents the systems submitted to the STD ALBAYZIN 2014 evaluation, held as a part of the ALBAYZIN 2014 evaluation campaign within the context of the IberSPEECH 2014 conference. This is the first STD evaluation that deals with Spanish language. The evaluation consists of retrieving the speech files that contain the search terms, indicating their start and end times within the appropriate speech file, along with a score value that reflects the confidence given to the detection of the search term. The evaluation is conducted on a Spanish spontaneous speech database, which comprises a set of talks from workshops and amounts to about 7 h of speech. We present the database, the evaluation metrics, the systems submitted to the evaluation, the results, and a detailed discussion. Four different research groups took part in the evaluation. Evaluation results show reasonable performance for moderate out-of-vocabulary term rate. This paper compares the systems submitted to the evaluation and makes a deep analysis based on some search term properties (term length, in-vocabulary/out-of-vocabulary terms, single-word/multi-word terms, and in-language/foreign terms).

Keywords: Spoken term detection; Spanish; International evaluation; Search on spontaneous speech

Introduction

The enormous amount of information stored in audio and audiovisual repositories promotes the development of efficient methods that aim at retrieving the stored information. For audio content search, significant research has been conducted in spoken document retrieval (SDR), keyword spotting, spoken term detection (STD), and query-by-example. Spoken term detection aims at finding a list of terms (composed of individual words or multiple words) within audio archives, and has been receiving much interest for years from the likes of IBM [1–3], BBN [4], SRI and OGI [5–7], BUT [8–10], Microsoft [11], QUT [12, 13], JHU [14–16], Fraunhofer IAIS/NTNU/TUD [17], NTU [18, 19], IDIAP [20], etc. In addition, several evaluations including SDR, STD, and query-by-example STD have been recently proposed [21–31].

Given the increasing interest in STD evaluations around the world, we organized an international evaluation of STD in the context of the ALBAYZIN 2014 evaluation campaign. This campaign is an internationally open set of evaluations supported by the Spanish Network of Speech Technologies (RTTH [32]) and the ISCA Special Interest Group on Iberian Languages (SIG-IL [33]), which have been held every 2 years since 2006. The evaluation campaigns provide an objective mechanism to compare different systems and are a powerful way to promote research

*Correspondence: javier.tejedor@depeca.uah.es
[1]GEINTRA, Universidad de Alcalá, Campus Universitario. Ctra. Madrid-Barcelona, km.33,600, Alcalá de Henares, Madrid, Spain
Full list of author information is available at the end of the article

on different speech technologies (e.g., speech segmentation [34], speaker diarization [35], language recognition [36], query-by-example spoken term detection [37], and speech synthesis [38] in the ALBAYZIN 2010 and 2012 evaluation campaigns). This year, this campaign has been held during the IberSPEECH 2014 conference [39].

Introduction to spoken term detection technology

Spoken term detection relies on a text-based input, commonly the orthographic transcription of the search term. Spoken term detection systems are typically composed of two main stages: indexing by an automatic speech recognition (ASR) subsystem, and then search by a detection subsystem, as depicted in Fig. 1. The ASR subsystem decodes the input speech signal in terms of word/subword lattices. The detection subsystem integrates a *term detector* and a *decision maker*. The term detector searches for putative detections of the terms in the word/subword lattices. The decision maker decides whether each detection is reliable enough to be considered as a hit or should be rejected as a false alarm (FA). Finally, a tool provided by the National Institute of Standards and Technology (NIST) is commonly used for performance evaluation [40].

There are two main approaches to STD: the word-based approach [6, 41–45] that searches for terms in the output of a large vocabulary continuous speech recognition (LVCSR) system, and the subword-based approach which searches for subword representations of search terms within the output of a subword speech recognition system. The word-based STD approach typically obtains better performance than the subword-based approach thanks to the lexical information it employs. However, the subword-based approach has the unique advantage that it can detect terms that consist of words that are not in the recognizer's vocabulary — out-of-vocabulary (OOV) terms — whereas the word-based approach can only detect in-vocabulary (INV) terms. Several subword unit types have been employed in the subword-based approach, including word fragments [46], particles [47, 48], acoustic words [49], graphones [6, 7], multigrams [9, 50], syllables [51–53], and graphemes [54], although phonemes are the most commonly used due to their simplicity and natural relationship with spoken languages [41, 55–59]. In order to exploit the relative advantages of the word and phoneme-based approaches, it has been proposed to combine these two approaches by using the word-based approach to detect INV terms and the subword-based approach to detect OOV terms, e.g., [41, 56, 60–64]. A hybrid approach that fuses word and subword lattices and then searches for both INV terms and OOV terms in the hybrid lattices has also been proposed [11, 65]. Another hybrid approach uses word/subword mixed lexica and language models to generate hybrid lattices [7, 10, 66]. A recent hybrid approach employs word confusion networks (WCNs) during ASR decoding and next incorporates a probabilistic phonetic retrieval (PPR) framework to deal with OOV terms [67]. Kaldi STD system [68–70] employs a word-based approach for term detection and a method based on proxy words (i.e., replacing the OOV term by the most similar in-vocabulary term/terms) to detect OOV terms [71].

Spoken term detection under the IARPA BABEL program and Open KWS

Significant research has been conducted on STD under the IARPA BABEL program [72]. This program was born in 2011 and aims at developing fully automatic and noise-robust speech recognition systems in limited time (e.g., 1 week) and with limited amount of transcribed training data, so that they can be applied to any language in order to process massive amounts of speech data recorded in challenging real-world situations. Spoken term detection perfectly fits within the scope of this program, which includes keyword search algorithms and low resource languages within its research areas. This

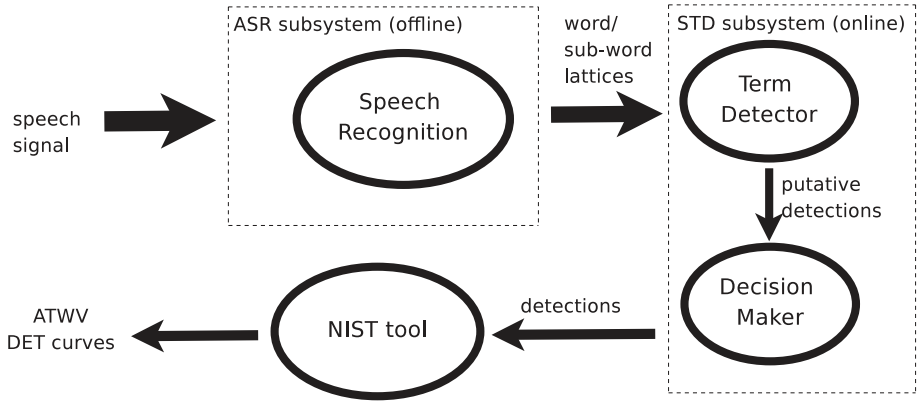

Fig. 1 Standard STD system architecture and evaluation. The standard spoken term detection system architecture and evaluation

program supports research in the following languages, corresponding to base period, option period 1, and option period 2 releases: Cantonese, Pashto, Tagalog, Turkish, Vietnamese, Assamese, Bengali, Haitian Creole, Lao, Zulu, Tamil, Kurmanji Kurdish, Tok Pisin, Cebuano, Kazakh, Telugu, Lithuanian, and Swahili. Since 2013, NIST has been organizing an annual open STD evaluation called NIST Open Keyword Search (KWS), which is closely related to the BABEL program but open to other research groups besides BABEL participants (more information in "Comparison to other evaluations" section). In this section, we will review some relevant results arisen from research in this framework, which focuses on OOV term detection, score normalization, and system combination.

The work presented in [73] focused on OOV term detection from different recognition units (word, syllable, and word fragment) and two search strategies (whole unit fuzzy search and phone fuzzy search) from the lattices obtained during the ASR process. For the phone fuzzy search, each recognition unit is first split into phones. Experimental results showed that (1) phone-based search outperformed the whole unit-based search for OOV terms, and whole-word search performed the best for INV terms; (2) the syllable models outperformed the word fragment models for the phone search; and (3) system combination from different recognition units and search strategies performed better than each individual system.

Wang and Metze [74] focused on score normalization and proposed a term-specific threshold that uses the confidence scores assigned to all the detections of the given term to compute the final score for each detection.

Karakos et al. [75] presented a new score normalization approach based on the combination of an unsupervised linear fit method and a supervised linear model method (Powell's method [76]) from several input features such as posterior probability, keyword length, false alarm probability, etc.

Chiu and Rudnicky [77] proposed a score normalization based on *word burst* (i.e., words of interest that occur near each other in the speech content) by penalizing the term detections that do not occur near other detections of the same term.

Deep neural networks (DNNs) as input for a Hidden Markov Model (HMM)-Gaussian Mixture Model (GMM) classifier have also shown their potential [78–81].

Language-independent and unsupervised training-based approaches have also been considered within this program aiming at building a system for an unknown language [82]. The limited data corresponding to some languages covered in the program (Cantonese, Pashto, Turkish, Tagalog, Vietnamese, Assamese, Bengali, Haitian Creole, Lao, and Zulu) were used for system training. The system is based on multi-lingual bottle-neck DNNs and Hidden Markov Model Toolkit (HTK) [83] for training and decoding and the IBM keyword search system for term detection [84]. Results showed that INV term performance is good for languages (e.g., Haitian Creole) whose phonetic structure is similar to that of the languages used for system training.

Various subword unit types (syllable, phone, grapheme, and automatically discovered) were investigated in [85] in the framework of lattice- and consensus network-based exact match term detection. Experimental results showed that (1) the automatically discovered units performed the best in isolation, (2) the combination of all the subword unit types for detection fusion significantly outperformed each subword unit type, and (3) fusion of the phone- and grapheme-based systems performed better than each individual system.

Lee et al. [86] investigated graph-based re-ranking techniques for scoring detection in STD systems for low-resource languages (Assamese, Bengali, and Lao). A node in the graph represents a hypothesized region of the given term, and connections are created from acoustically similar hypothesized regions. The STD system is based on fuzzy matching and different word/subword units (word, syllable, morpheme, and phoneme).

Ma et al. [87] proposed a combined approach for detection re-scoring from linear interpolation of a rule-based detection re-scoring system, a logistic regression-based detection re-scoring system, and a rank learning-based detection re-scoring system. The detection re-scoring system based on word-burst features (e.g., number, strength, and proximity of neighbor hypothesis, etc.), consensus network features (e.g., posterior probability, number of hit arcs, number of average arcs per bin, etc.), and acoustic features (e.g., pitch, number of unvoiced frames, jitter, etc.).

Chiu et al. [88] proposed combining finite state transducer- and confusion network-based STD systems from DNN, bottle-neck, and perceptual linear prediction (PLP) acoustic features.

A novel two-stage discriminative score normalization method was presented in [89]. The term detector employed word lattices obtained from an LVCSR system to output term detections. Next, the discriminative score normalization method relies on a multi-layer perceptron (MLP)-based confidence measure from two novel features. These novel features are the ranking score, computed as the rank of the posterior probability of the detection compared to the posterior probability of all the arcs in the lattice where the detection resides and the relative posterior probability of the detection compared to the maximum posterior probability within the arcs in the lattice where the detection resides. The new confidence score is then taken by an ATWV-oriented score normalization in the second stage, which optimizes the final score for the evaluation metric.

Wegmann et al. [90] presented a system where detections of several ASR systems were combined. ASR systems were built from HTK [83] and Kaldi [68] tools and employed PLP and bottle-neck acoustic features. More interestingly, this work also made an analysis of the ATWV performance from different approaches. The first approach consisted on setting the optimal threshold for each term from the ground-truth information. This analysis showed that there are important performance gaps in ATWV due to the thresholding algorithm employed, suggesting that a better threshold selection will produce significant performance gains. The second approach is based on bootstrapping techniques to show the ATWV results of randomly selected groups of terms. The different distribution of the ATWV performance across the different term groups showed that ATWV heavily depends on the selected terms, and even that small changes in the ASR system accuracy can cause large changes in the STD performance.

Several score normalization and system combination approaches were presented in [91]. Score normalization based on term-dependent thresholding, rank normalization and mapping back to posteriors, sum-to-one normalization, and machine learning. The term-dependent thresholding simply re-scores the detection by considering the confidence scores of all the detections of the given term in the ATWV formulation. The rank normalization is based on the false alarm rate for the given term as score normalization value for each term detection. The mapping back to posteriors approach relies on the average posteriors of the detections of all the terms except that being detected that are ranked in the same position within the detection list for the given term. The sum-to-one approach normalizes the score of the detection by the sum of all the scores of the detections of the given term. The machine learning approach is based on a linear model by applying the Powell's method [76] to maximize ATWV performance from several input features (e.g., rank normalization, mapping back to posteriors, term length, etc.). System combination merged the detections from different STD systems that rely on different approaches (e.g., GMM-based and DNN-based HMMs) and combined the detection scores from Powell's method.

Su et al. [53] proposed syllable-weighted finite state transducer (WFST) for speech indexing and direct search on syllable- and word-WFST for term detection. The word-WFST is obtained by syllable-to-word mapping from the original syllable-WFST. Experiments showed that the system combination from word- and syllable-WFST at detection level significantly outperforms each individual system.

Chen et al. [92] presented a novel subword unit-based approach that focused on pronunciation prediction. To get the optimal set of subword units, the pronunciation prediction is first based on syllables, which are then converted to a more specific subword units (similar to morphemes), according to a certain lexicon segmentation that obtains the highest language model score for each pronunciation in the lexicon. For OOV term detection, the phoneme transcription of the terms is obtained with the sequitur grapheme-to-phoneme tool [93], which is next mapped to subword units. The novel subword approach outperformed the system performance of word-, syllable-, and phoneme-based units. In addition, system combination from word, novel subword, syllable, and phoneme units showed significant performance gains over each individual system.

Trmal et al. [94] proposed system combination from different ASR systems that employ different configurations in terms of acoustic features and acoustic models (e.g., subspace GMMs (SGMMs), DNNs, and bottle-neck features). Kaldi STD system [68–70] was used for term detection in all the systems. A syllable-based lexicon expansion was proposed for OOV keyword search. Point process models (PPMs) were also employed for keyword search. These are based on whole-word, event-based acoustic modeling and phone-based search [95, 96]. Since they are phone based, OOV term detection is not an issue for the PPM-based STD systems. Experimental results showed that (1) the combination of PPM-based STD and Kaldi-based STD effectively improved the STD performance, and (2) the lexicon expansion generally outperforms the system performance.

Keyword spotting under the DARPA RATS program

The DARPA Robust Automatic Transcription of Speech (RATS) program also includes keyword spotting within its research areas. Different to the BABEL program, DARPA RATS program mainly focuses on speech recognition under highly noisy communication channels, where typically speech signals of less than 10 dB are specified. Two main languages have been employed in this program: Levantine Arabic and Farsi. For these languages, significant research has also been carried out in keyword spotting. In this section, we will try to summarize the most significant research in this program, which mainly focuses on score normalization and system combination.

A keyword spotting system was presented in [97] with a score normalization approach based on the false alarm probability of the given term. In addition, a *white list*-based approach in the ASR system was also presented. This approach modifies the beam pruning produced at recognition, by keeping alive (using a wider beam) those states that form a detection of a term in the white list. Since the white list contains all the search terms, all the term detections are very unlikely to be pruned.

The system presented in [98] used also the white list approach presented in [97] and focused on system combination from word lattices and phone confusion networks. Both word lattices and phone confusion networks were generated from different ASR systems that employed different configurations (Mel-frequency cepstral coefficient (MFCC), PLP, GMM, SGMM, etc.). Detections of the different ASR systems were combined using logistic regression.

Deep neural networks have also been employed for developing keyword spotting systems under the DARPA RATS program [99]. In this work, several word- and subword-based systems were combined with the system combination approach presented in [91]. A similar work based on DNNs, GMMs, and convolutional neural networks (CNNs) for acoustic modeling and various signal processing features (standard cepstral and filter-bank features, noise-robust features, and MLP features) was presented in [100]. This employs word- and phone-based ASR systems to produce a set of term detections that are next fused with the logistic regression-based approach presented in [98].

Mangu et al. [101] employed CNNs, DNNs, and GMMs as acoustic modeling, audio segmentation based on GMMs and DNNs, word lattices as ASR output, phone-WFST for keyword search, and system combination. System combination took the output of the different ASR systems and merged the detections of all the systems. Detection scores are normalized by the sum of all the scores of the detections of the given term. Experimental results showed that (1) CNNs perform very well for keyword search, (2) audio segmentation plays a very important role in keyword search, and (3) system combination yields significant performance gains.

Seigel et al. [102] employed a system combination approach based on word and grapheme ASR. Word- and grapheme-based lattices are first produced and then used for term search. Conditional random field (CRF) models are used for detection scoring in a discriminative confidence scoring framework. The input features to the CRF are related to the lattice information, contextual posterior features, and unigram prior features.

Mitra et al. [103] focused on system combination from word lattices. The word lattices are obtained from different GMM-HMM speech recognition systems that employ different sets of acoustic features (e.g., PLP, normalized modulation cepstral coefficients, and modulation of medium duration speech amplitude), along with various feature combination and dimensionality reduction techniques (principal component analysis, heteroscedastic linear discriminant analysis, and nonlinear autoencoder network). Experiments showed that the feature combination (prior combination) and the detection combination from individual ASR systems (posterior combination) yield significant performance gains.

The rest of the paper is organized as follows: the next section presents the STD evaluation and includes an evaluation description, the metric used, the database released for experimentation, a comparison with previous evaluations, and the participants involved in the evaluation. Next, we present the different systems submitted to the evaluation. Results along with discussion are presented in a separate section, and finally conclusions are presented.

Spoken term detection evaluation

STD evaluation overview

This evaluation involves searching a list of terms within speech content. Therefore, the evaluation is designed for research groups working on speech indexing and retrieval and speech recognition as well. In other words, the STD evaluation focuses on retrieving the appropriate audio files, with the occurrences and timestamps, which contain any of those terms.

The evaluation consists of searching a training/development term list within training/development speech data and searching a test term list within test speech data. The evaluation result ranking is based on the system performance when searching the test terms within test speech data. Participants can use the training/development data for system training and tuning, but any additional data can also be employed.

Participants could submit a primary system and up to 4 contrastive systems. No manual intervention is allowed for each system developed to generate the final output file, and hence all the developed systems must be fully automatic. Listening to the test data or any other human interaction with the test data is forbidden before all the evaluation results in terms of the performance of the systems in test data (i.e., evaluation result ranking) have been sent back to the participants. The standard XML-based format corresponding to the NIST STD 2006 evaluation [22] has been used for building the system output file.

Evaluation metric

In STD, a hypothesized occurrence is called a *detection*; if the detection corresponds to an actual occurrence, it is called a *hit*, otherwise it is a *false alarm*. If an actual occurrence is not detected, this is called a *miss*. The ATWV proposed by NIST [22] has been used as the main metric for the evaluation. This metric integrates the hit rate and false alarm rate of each term into a single metric and then averages over all the terms:

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

where Δ denotes the set of terms and $|\Delta|$ is the number of terms in this set. N_{hit}^{K} and N_{FA}^{K} represent the numbers of hits and false alarms of term K, respectively, and N_{true}^{K} is the number of actual occurrences of K in the audio. T denotes the audio length in seconds, and β is a weight factor set to 999.9, as in the ATWV proposed by NIST [4]. This weight factor causes an emphasis placed on recall compared to precision in the ratio 10:1.

ATWV represents the TWV for the threshold set by the STD system (usually tuned on development data). An additional metric, called maximum term weighted value (MTWV) [22] can also be used to evaluate the performance of an STD system. This MTWV is the maximum TWV achieved by the STD system for all possible thresholds and hence does not depend on the tuned threshold. Therefore, this MTWV represents an upper-bound of the performance obtained by the STD system. Results based on this metric are also presented to evaluate the system performance with respect to threshold selection.

In addition to ATWV and MTWV, NIST also proposed a detection error tradeoff (DET) curve [104] to evaluate the performance of an STD system working at various miss/FA ratios. Although DET curves were not used for the evaluation itself, they are also presented in this paper for system comparison.

The NIST STD evaluation tool [40] was employed to compute MTWV, ATWV, and DET curves.

Additionally, precision, recall, and F-measure values are also presented in this paper to evaluate system performance. Whereas the original ATWV metric proposed by NIST gives more emphasis to recall than to precision (in other words, it is more important a miss than a false alarm), F-measure assigns the same cost to precision and recall values. Therefore, F-measure allows us to compare the system performance in a different way. However, it must be noted that the systems submitted to the evaluation were tuned and optimized towards ATWV.

Database

The database used for the evaluation consists of a set of talks extracted from the MAVIR workshops [105] held in 2006, 2007, and 2008 (corpus MAVIR 2006, 2007, and 2008) that contain speakers from Spain and Latin America (henceforth MAVIR corpus or database). The MAVIR corpus contains 3 recordings in English and 10 recordings in Spanish, but only the recordings in Spanish were used for the evaluation.

The MAVIR Spanish data consist of spontaneous speech files, each containing different speakers, which amount to about 7 h of speech and are further divided for the purpose of this evaluation into training/development and test sets. There are 20 male and 3 female speakers in the MAVIR Spanish database. The data were also manually annotated in an orthographic form, but timestamps were only set for phrase boundaries. To prepare the data for the evaluation, we manually added the timestamps for the roughly 6000 occurrences of spoken terms used in the training/development and test evaluation sets.

The speech data were originally recorded in several audio formats (PCM mono and stereo, MP3, 22.05 KHz., 48 KHz., etc.). All data were converted to PCM, 16 KHz., single channel, 16 bits per sample using SoX tool [106]. Recordings were made with the same equipment, a Digital TASCAM DAT model DA-P1, except for one recording. Different microphones were used for the different recordings. They mainly consisted of tabletop or floor standing microphones, but in one case a lavalier microphone was used. The distance from the mouth of the speaker to the microphone varies and was not particularly controlled, but in most cases the distance was smaller than 50 cm. All the speech contain real and spontaneous speech of MAVIR workshops in a real setting. Thus, the recordings were made in large conference rooms with capacity for over a hundred people and a large amount of people in the conference room. This poses additional challenges including background noise (particularly babble noise) and reverberation. The realistic settings and the different nature of the spontaneous speech in this database make it appealing and challenging enough for our evaluation. Table 1 includes some database features such as the number of word occurrences, duration, and signal-to-noise ratio (SNR) [107] of each speech file in the MAVIR Spanish database.

Training/development data amount to about 5 h of speech extracted from 7 out of the 10 speech files of the MAVIR Spanish database and contain 15 male and 2 female speakers. This material was made available to the participants including the orthographic transcription and the timestamps for phrase boundaries [108]. However, there is no constraint in the amount of training/development data beyond the MAVIR corpus that can be employed to build the systems. The training/development term list consists of 346 terms. Each term can be composed of a single word or multiple words and its length varies between 5 and 25 phonemes. Ground truth labels and evaluation tools were provided to the participants by the date of the release. There are 4192 occurrences of those terms in the training/development data. Table 2 includes information related to the training/development term list, and Fig. 2 shows the histogram with the number of terms that contain a certain number of phonemes.

Test data amount to about 2 h of speech extracted from the other 3 speech files of the MAVIR Spanish database not used as training/development data and contain 5 male and 1 female speakers. The test term list consists of 202 terms. Each term can be composed of one or multiple words and its length varies between 5

Table 1 MAVIR database characteristics. "train/dev" stands for training/development, "occ." stands for occurrences, "min" stands for minutes, "SNR" for signal-to-noise ratio, and "dB" for decibels

File ID	Dataset	# word occ.	Duration (min)	# speakers	SNR
MAVIR-02	train/dev	13432	74.51	7 (7 male)	2.1 dB
MAVIR-03	train/dev	6681	38.18	2 (1 male, 1 female)	15.8 dB
MAVIR-06	train/dev	4332	29.15	3 (2 males, 1 female)	12.0 dB
MAVIR-07	train/dev	3831	21.78	2 (2 males)	10.6 dB
MAVIR-08	train/dev	3356	18.90	1 (1 male)	7.5 dB
MAVIR-09	train/dev	11179	70.05	1 (1 male)	12.3 dB
MAVIR-12	train/dev	11168	67.66	1 (1 male)	11.1 dB
MAVIR-04	test	9310	57.36	4 (3 males, 1 female)	10.2 dB
MAVIR-11	test	3130	20.33	1 (1 male)	9.2 dB
MAVIR-13	test	7837	43.61	1 (1 male)	11.1 dB
ALL	train/dev	53979	320.23	17 (15 males and 2 females)	-
ALL	test	20277	121.3	6 (5 males and 1 female)	-

and 23 phonemes. No ground truth labels corresponding to the test data were given to the participants until the organizers have sent them back the evaluation results. There are 2054 occurrences of the test terms in the test data. Table 3 includes information related to the test term list, and Fig. 3 shows the histogram with the number of terms that contain a certain number of phonemes.

Table 2 Twenty most and less occurrence terms of the training/development term list along with the number of phonemes (# phonemes) and the number of occurrences in the training/development data

Term (# phonemes)	# occurrences	Term (# phonemes)	# occurrences
Información (11)	153	mercurio (8)	1
También (7)	113	música pop (9)	1
Mercado (7)	111	mystic (6)	1
Internet (8)	74	nostradamus (11)	1
Empresas (8)	71	pacífico (8)	1
Importante (10)	69	patagonia (9)	1
Ustedes (7)	66	real alcázar (11)	1
Investigación (13)	52	reino unido (10)	1
Momento (7)	50	repsol (6)	1
Imágenes (8)	46	salamanca (9)	1
Simplemente (11)	46	sudamérica (10)	1
General (7)	43	suecia (6)	1
Motores (7)	43	taiwán (6)	1
Noventa (7)	39	torres quevedo (11)	1
Primero (7)	39	valencia (8)	1
Después (7)	38	venezuela (9)	1
Utilizar (8)	38	verity (6)	1
Siempre (7)	36	windows mobile (13)	1
Trabajo (7)	36	zaragoza (8)	1
Proyectos (9)	34	zurich (5)	1

Table 4 includes some information related to the training/development and test data files used in the evaluation such as the number of term occurrences and the average number of term occurrences per minute, the number of different terms, and the average number of different terms and their occurrences per minute. It must be noted that, although the length of the speech data used for training/development is greater than that of the test data, the average number of occurrences per minute in both sets is similar.

All the terms selected for both sets (training/ development and test) aimed at building a realistic scenario for STD, by including high occurrence terms, low occurrence terms, foreign terms, single-word and multi-word terms, in-vocabulary and out-of-vocabulary terms, and different length terms. Each training/development term has one or more occurrences in the training/development speech data and each test term has one or more occurrences in the test speech data. Table 5 includes some features of the training/development and test term lists such as the number of in-language and foreign terms, the number of single-word and multi-word terms, and the number of in-vocabulary and out-of-vocabulary terms, along with the number of occurrences of each set in the speech data. A term is considered OOV if this does not appear in the training/development speech data provided by the organizers. It must be noted that a multi-word term is considered OOV in case any of the words that form the term is OOV. Therefore, in case the OOV terms appear in the vocabulary of the ASR systems, these have been added to the ASR system vocabulary from other sources (web, newspapers, other speech databases, etc.).

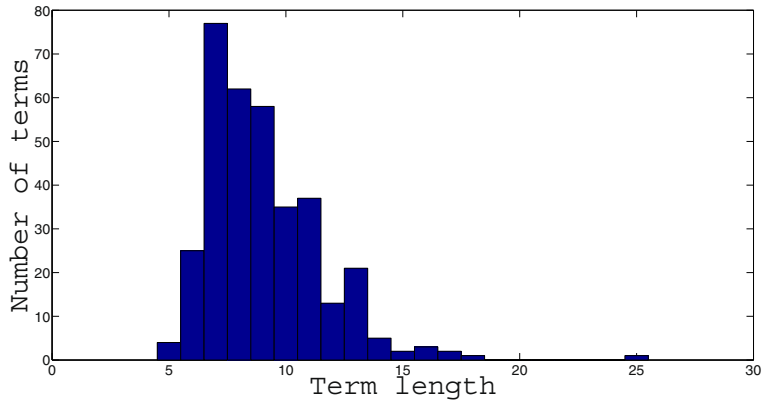

Fig. 2 Histogram of the training/development term list. Histogram of the number of terms in the training/development term list with respect to the term length (in phonemes)

Comparison to other evaluations

In the last years, several evaluations in the field of spoken term detection have taken place. In this section, we review the former evaluations mainly to highlight the differences with the evaluation presented in this paper. The National Institute of Standards and Technology of the USA organized in 2006 the NIST STD evaluation [22]. The data contained speech in English, Mandarin Chinese, and Modern Standard and Levantine Arabic. In this evaluation, the nature of the speech included conversational telephone speech (CTS), broadcast news (BNews) speech, and speech recorded in roundtable meeting rooms (RTMeet) with distantly placed microphones (this last type is used only for English). Of the three different types of speech, the last one is the most similar to the nature of the speech in our evaluation, although there are some differences in terms of the size of the room, larger in our case, which has a negative impact on the system performance due to the reverberation; also, the use of amplification of the audio in the conference rooms is not present in the case of a roundtable meeting. The NIST STD 2006 evaluation results are publicly available [109], and are a very interesting result to analyze the influence of the language and the nature of speech on STD results. Table 6 presents the best results obtained by the evaluation participants for each condition. With respect to the type of speech, it is clear from Table 6 that results using microphone speech, particularly distant microphones in less controlled settings than in audiovisual studios (such as in broadcast news) or close-talking conversational telephone data, are definitely much more limited. With respect to the language, English is the language with more resources and for which more research has been done so far. When applying similar technology to languages for which less specific research has been conducted, performance decreases are observed.

NIST has also carried out a new set of evaluations called NIST Open KWS in the last years [30, 31]. These evaluations, named Open KWS 2013 and Open KWS 2014, are very similar to the former NIST STD 2006 evaluation. These were integrated within the BABEL program and aimed at building STD systems in a limited time for low-resource languages (Vietnamese and Tamil). These new evaluations were only conducted on CTS data on a

Table 3 Twenty most and less occurrence terms of the test term list along with the number of phonemes (# phonemes) and the number of occurrences in the test data

Term (# phonemes)	# occurrences	Term (# phonemes)	# occurrences
También (7)	93	flebitis (8)	1
Información (11)	92	gómez (5)	1
Ejemplo (7)	54	iberoamericanas (15)	1
Sistemas (8)	46	infecciones urinarias (19)	1
Respuestas (10)	44	latinoamericano (15)	1
Usuarios (8)	40	luis rodrigo (11)	1
Trabajar (8)	39	marcin (6)	1
Hipótesis (8)	38	nueva gales del sur (16)	1
Entonces (8)	37	open directory (13)	1
Implicación (11)	31	pablo serrano (11)	1
Nosotros (8)	30	paz iglesias (11)	1
Entidades (9)	28	península ibérica (16)	1
Idiomas (7)	27	pepsi twist (10)	1
Imágenes (8)	26	potter (5)	1
Bastante (8)	22	reino unido (10)	1
Distintos (9)	21	río tajo (7)	1
Resultados (10)	21	sinamed (7)	1
Encontrar (9)	19	strathclyde (10)	1
Formularios (11)	19	víctor (6)	1
Importante (10)	19	webometrics (10)	1

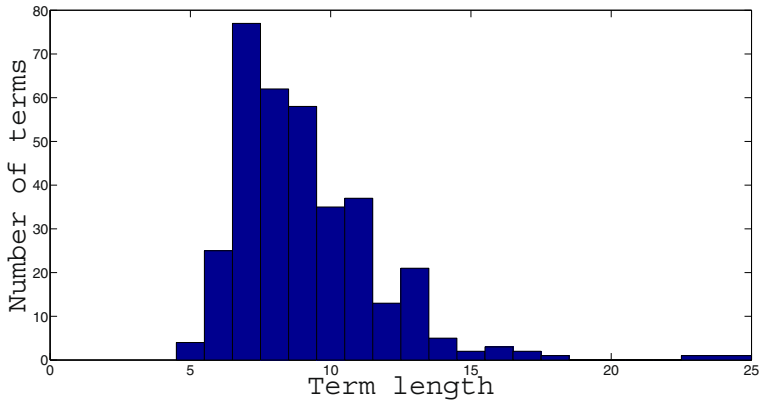

Fig. 3 Histogram of the test term list. Histogram of the number of terms in the test term list with respect to the term length (in phonemes)

surprise language that was announced only a few (4 or less) weeks before the evaluation. Best performance in the NIST Open KWS 2013 evaluation is ATWV=0.6248 [110] under the Full Language Pack (FullLP) condition, for which 20 h of word-transcribed scripted speech, 80 h of word-transcribed CTS, and a pronunciation lexicon were given to participants. In the works describing systems on the surprise language (i.e., Tamil) of the Open KWS 2014 evaluation [53, 92, 94, 111–117], ATWV=0.5802 is the best performance obtained under the FullLP condition, for which 60 h of transcribed speech and a pronunciation lexicon were given to participants.

In our evaluation, the audio contains microphone recordings of real talks in real workshops, in large conference rooms with public. Microphones, conference rooms, and even recording conditions change from one recording to another. Microphones are not close-talking microphones but mainly tabletop and ground standing microphones. This difference in the evaluation conditions makes our evaluation pose different challenges and makes it difficult to compare the results obtained in our evaluation to those of the previous NIST STD evaluations.

Additionally, a new round of STD evaluations has been organized within NTCIR conferences [23, 28, 29]. Data used in these evaluations contained spontaneous speech in Japanese provided by the National Institute for Japanese language and spontaneous speech recorded during seven editions of the Spoken Document Processing Workshop. These evaluations also provide the manual transcription of the speech data and the output of an LVCSR system to the participants. Table 7 presents the best result obtained in each individual evaluation, where the *F*-measure was used as the evaluation metric. Although our evaluation could be similar in terms of speech nature to these NTCIR STD evaluations (speech recorded in real workshops), we do not provide any kind of information apart from the speech content, the list of terms, and the training/development ground-truth files to the participants.

Table 4 MAVIR training/development and test data file characteristics. "train/dev" stands for training/development, "occ." stands for occurrences, and "min" stands for minutes

File ID	Dataset	# occ.	# occ./min	# different terms	# different terms/min
MAVIR-02	train/dev	1016	13.6	203	2.7
MAVIR-03	train/dev	653	17.1	153	4.0
MAVIR-06	train/dev	446	15.3	126	4.3
MAVIR-07	train/dev	296	13.6	104	4.8
MAVIR-08	train/dev	200	10.6	76	4.0
MAVIR-09	train/dev	910	13.0	199	2.8
MAVIR-12	train/dev	671	9.9	129	1.9
MAVIR-04	test	1026	17.9	167	2.9
MAVIR-11	test	414	20.4	70	3.4
MAVIR-13	test	614	14.1	98	2.2
ALL	train/dev	4192	13.1	346	1.1
ALL	test	2054	16.9	202	1.7

Table 5 Training/development and test term list characteristics. "train/dev" stands for training/development, "IN-LANG" refers to in-language terms, "OUT-LANG" to foreign terms, "SINGLE" to single-word terms, "MULTI" to multi-word terms, "INV" to in-vocabulary terms, "OOV" to out-of-vocabulary terms, and "occ." stands for occurrences

Term list	# IN-LANG/OUT-LANG terms (occ.)	# SINGLE/MULTI terms (occ.)	# INV/OOV terms (occ.)
Train/dev	330 (4061)/16 (131)	325 (4166)/21 (26)	346 (4192)/0 (0)
Test	189 (2020)/13 (34)	185 (2032)/17 (22)	150 (1840)/52 (214)

In addition, our evaluation makes use of other language, employs a larger list of terms, and defines disjoint training/development and test term lists to measure the generalization capability of the systems. The evaluation presented here is, to the best of our knowledge, the first STD evaluation that deals with the Spanish language.

Participants

Six different systems were submitted from four different research groups to the spoken term detection ALBAYZIN 2014 evaluation. Participants are listed in Table 8. About 3 months were given to the participants for system development, and hence the STD evaluation focuses on building STD systems in limited time. The training/development and test data were released to the participants in different periods. Training/development data (i.e., training/development speech data, training/development term list, training/development ground-truth labels, orthographic transcription and timestamps for phrase boundaries in the training/development speech data, and evaluation tools) were released at the end of June 2014. The test data (i.e., test speech data and test term list) were released at the beginning of September 2014. The final system submission was due at the end of September 2014. Final results were discussed at IberSPEECH 2014 conference at the end of November 2014.

Additional considerations for the STD evaluation design

The first STD evaluation, which was born in 2006 and was held by NIST [22], aimed at finding a set of terms within huge audio archives. In 2000, Garofolo claimed that the information extraction in large audio repositories was a solved problem by means of the LVCSR systems [118]. In this way, a search of the terms of interest within their output would be enough for practical applications. However, these LVCSR systems suffer from the OOV problem, since OOV terms are impossible to retrieve by standard LVCSR systems. In addition, Logan showed that about 10 % of the user queries to a spoken information retrieval system contain OOV terms [119]. Therefore, it is reasonable that STD evaluations focus on OOV term detection. Our evaluation also considers the OOV term detection in a great extent, by incorporating some terms that do not appear in the training/development speech data to the test term list. The systems need to deal with these OOV terms so that the final performance is not degraded. On the one hand, they can incorporate these OOV terms to their LVCSR system vocabulary, in case significant amount of training/development material is obtained from external sources (e.g., web, newspapers, broadcast news, etc.). Otherwise, systems must rely on some other approach (e.g., subword-unit ASR) to retrieve them. Moreover, by incorporating INV and OOV terms in the test term list, organizers greatly encouraged the participants to build hybrid systems from the combination of word- and subword unit-based STD systems. Since both phone and syllable sets in Spanish language are well defined [120], these two types of subword units can effectively deal with the OOV terms.

The MAVIR database chosen for the evaluation consists of highly spontaneous speech from real workshops. Given this database condition, there is an inherent difficulty for term detection. In addition, STD and, in general, ASR systems significantly degrade their performance when training/development data belong to a different

Table 6 Best performance (in terms of actual term weighted value, ATWV) obtained by the different participants of the NIST STD 2006 evaluation in the different conditions: "CTS" stands for conversational telephone speech, "BNews" for broadcast news, and "RTMeet" for speech recorded in roundtable meeting rooms

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

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

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

Table 8 Participants in the spoken term detection ALBAYZIN 2014 evaluation along with the systems submitted

Team ID	Research institution	Systems
GTM	AtlantTIC Research Center, University of Vigo, Spain	Fusion, WL-Kaldi
GTH	University Politécnica of Madrid, Spain	W1B-HTK
ATVS-GEINTRA	University Autónoma of Madrid - University of Alcalá, Spain	WL-ATWV-Kaldi, WL-WER-Kaldi
VivoLab	University of Zaragoza, Spain	P1B-HTK

domain or pose different acoustic conditions to those of the test data. To alleviate this problem in the STD evaluation, organizers paid special attention in preparing limited training/development data from the same domain and with *similar* acoustic conditions (microphone speech from workshops) to that of the test data. This material could be used by the participants to train and tune their systems (see "Database" section). However, it must be noted that different microphones were used for each recorded file in the MAVIR database, and hence the acoustic conditions slightly vary from one file to another. Moreover, the limited training/development data, which amount to 5 h of speech, adds another challenge to the evaluation, aiming at building STD systems with limited data that match the test data conditions.

Systems

In this section, we describe the systems submitted to the STD evaluation. The systems appear in the same order that they are ranked in Tables 9–17. A summary of the systems is presented in Table 9.

Fusion-based STD system (fusion)

This system consists of the fusion of two different LVCSR-based STD systems, as depicted in Fig. 4; specifically, Kaldi-based and UVigo LVCSR-based STD systems were developed, which are described next.

Kaldi-based STD system

The Kaldi-based STD system comprises two different subsystems, as depicted in Fig. 5: An ASR subsystem is used to decode the speech utterances into word lattices; an STD subsystem integrates a term detector that searches for the input terms within the word lattices and a decision maker that ascertains reliable detections.

The ASR subsystem employs the Kaldi speech recognizer [68] to obtain word lattices from the input waveforms. Thirteen-dimensional PLP coefficients were used as acoustic features, and a state-of-the-art maximum likelihood (ML) acoustic model training strategy was employed. This training starts with a flat-start initialization of context-independent phonetic HMMs and ends with a speaker adaptive training (SAT) of state-clustered triphone HMMs with GMM output densities. After the ML-based acoustic model training stage, a universal background model (UBM) is built from speaker-transformed training data, which is next used to train an SGMM employed in the decoding stage to generate the word lattices.

The aforementioned acoustic models were trained using the Spanish data from 2006 TC-STAR automatic speech recognition evaluation campaign [121]. Specifically, the training data from the European Parliamentary Plenary Sessions (EPPS) and the Spanish Parliament Sessions, which were manually transcribed, were used for acoustic model training [122]. All the non-speech parts, the speech parts corresponding to transcriptions with pronunciation errors, incomplete sentences, and short speech utterances from the speech data were discarded. The training data amount to about 79 h of speech.

The language model (LM) was trained using a text database of 160 million words extracted from several sources: transcriptions of the European and Spanish Parliaments of the TC-STAR database, subtitles, books, newspapers, online courses, and the transcriptions of the MAVIR sessions included in the training/development data provided by the organizers [123]. For development experiments, a different LM was created for each MAVIR session, using the transcription of the session to obtain the optimum mixture of the partial LMs. The LM and the corresponding vocabulary created from all the training/development data files except one were then used to compute the detections of that file in a leave-one-out strategy. For the test data, the LM was generated using a normalized average of the weights obtained from the development sessions. It must be noted that the vocabulary was selected at the last stage of the LM training, once the partial LMs and their weights were computed. A trigram word LM trained with a vocabulary of 60k words and a Kneser-Ney discount strategy was used for the ASR subsystem.

The STD subsystem integrates the Kaldi term detector [68–70], which searches for the input terms within the word lattices obtained in the previous step. The lattice indexing technique, described in [124], first converts the word lattices of all the utterances in the speech data from

Table 9 System summary in terms of the ASR subsystem and STD subsystem employed. "prob." stands for probability

System ID	ASR subsystem	STD subsystem
Fusion	word lattices, word N-best	Fusion: Kaldi/search in word N-best term detector + posterior prob.-based decision maker
WL-Kaldi	word lattices	Kaldi term detector + posterior prob.-based decision maker
W1B-HTK	word 1-best	search in word 1-best + log likelihood-based decision maker
WL-ATWV-Kaldi	word lattices	Kaldi term detector + Kaldi decision maker
WL-WER-Kaldi	word lattices	Kaldi term detector + Kaldi decision maker
P1B-HTK	phone 1-best, phone lattices	search in phone 1-best + fusion-based decision maker

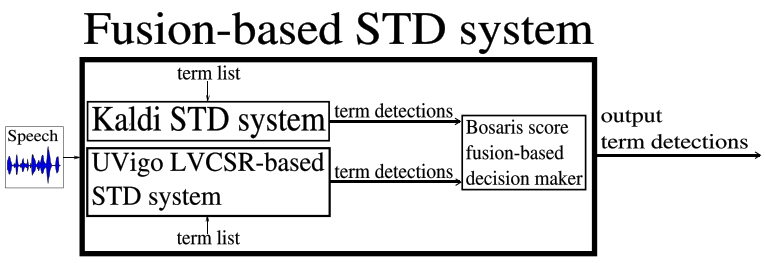

Fig. 4 Architecture of the fusion system

individual WFSTs to a single generalized factor transducer structure that stores the start-time, end-time, and the lattice posterior probability of each word token as a three-dimensional cost. This factor transducer represents an inverted index of all the word sequences contained in the lattices. Thus, given a search term, a simple finite state machine that accepts the term is created and composed with the factor transducer in order to obtain all the occurrences of the term in the speech data. The posterior probabilities of the lattice corresponding to all the words of the search term are accumulated, assigning a confidence score to each detection. The decision maker simply removes those detections with a confidence score below a predefined threshold.

The Kaldi spoken term detection system [68–70] handles OOV term search by means of a method called *proxy words* [71]. This method essentially consists of substituting each OOV word of the search term with acoustically similar INV proxy words, getting rid of the need of a subword-based system for OOV term search. However, this method was not used within this system, causing those terms containing any OOV words not to be detected at all.

The entire Kaldi-based STD system (both ASR and STD subsystems) was run on training/development data for parameter tuning, with the leave-one-out strategy explained before for LM building in the ASR subsystem. Next, the optimal parameters were used to hypothesize the detections corresponding to the test data.

UVigo LVCSR-based STD system

The UVigo LVCSR-based STD system is composed of two different subsystems, as depicted in Fig. 6: the ASR subsystem, which employs the UVigo LVCSR system [122], is used to decode the speech utterances in terms of word N-best hypotheses; the STD subsystem integrates a term detector that first obtains word meshes from the N-best hypotheses and then searches for the term within these word meshes, and a decision maker that ascertains reliable detections.

The UVigo LVCSR system comprises three different stages: audio segmentation, acoustic model selection, and ASR decoding. Thirteen MFCCs augmented with their delta and acceleration coefficients were used as acoustic features.

On the first stage, the audio is segmented (i.e., the input speech files are divided into more manageable speech segments). This segmentation is carried out by combining the output of the speaker segmentation strategy described in [125] with an energy-based voice activity detector (VAD). The speaker segmentation strategy divides the input audio signal into speech and non-speech segments, and speech segments are also divided into shorter segments according to their speaker. The energy-based VAD is then applied on these speaker-homogeneous segments in order to detect silence intervals, dividing each speaker turn into shorter chunks.

On the second stage, acoustic model selection is performed. Given a speech segment, the UVigo decoder [126] is used to conduct phone recognition using different sets of acoustic models to obtain a likelihood for each set. The set of acoustic models that obtained the highest likelihood was chosen, as it was considered to be the most suitable for decoding the corresponding speech segment. The set of acoustic models consists of 24 two-state demiphone acoustic models. Fourteen of the acoustic models were trained using the TC-STAR data described before, and they consist of gender-independent and gender-dependent acoustic models obtained from different partitions of the data. The remaining acoustic models were

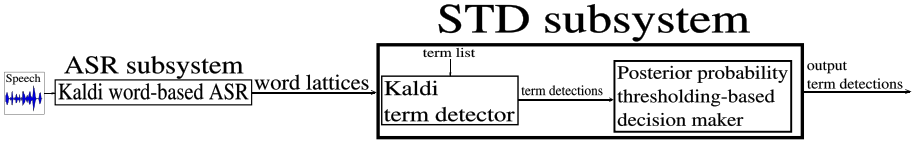

Fig. 5 Architecture of the Kaldi-based STD system of the fusion system

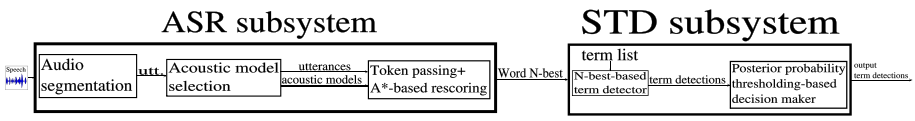

Fig. 6 Architecture of the UVigo LVCSR-based STD system of the fusion system. "utt." stands for utterance

obtained by adapting the TC-STAR models from different combinations of MAVIR training/development speech data.

The third stage employs the UVigo decoder to extract N-Best hypotheses for each speech segment; this decoder uses the token-passing algorithm with language model look-ahead and an A* stack strategy-based N-Best rescoring [126]. It must be noted that the acoustic model used for decoding is that selected in the acoustic model selection stage. As language model, the system employs the same as the Kaldi-based STD system.

The STD subsystem takes the word N-Best hypotheses produced by the LVCSR system as input for the lattice tool of the SRI-LM toolkit [127] and converts them to word meshes with posterior probabilities. Next, a search of the given term within the word mesh produces the term detections and outputs a posterior probability as score for each detection. The decision maker simply removes those detections whose posterior probability remains below a predefined threshold.

It must be noted that terms that do not appear in the LVCSR system vocabulary cannot be detected with this system.

As in the Kaldi-based STD system, the entire system was run on training/development data for parameter tuning. The optimal parameter set is next applied on test data.

System fusion

System fusion combines the output of the two systems described above to produce a more discriminative and better-calibrated score for each detection, aiming at taking advantage of the strengths of the individual approaches [128]. First, the optimal operating point was calculated on training/development data and applied to each individual system. After this, a global minimum zero-mean and unit-variance normalization was applied to prevent the scores of the individual systems to be in different ranges and to obtain term-independent scores. Finally, Bosaris toolkit [129] was used to construct a fusion scheme based on logistic regression; this procedure results in a new score for each detection, which is used by the decision maker of this system to output the final detections. The overlapping detections, i.e., detections of different terms on the same time interval, were removed by keeping the search term with the highest score.

Word lattice-based Kaldi STD system (WL-Kaldi)

This system is the Kaldi-based STD system described in the fusion system.

Word 1-best-based HTK STD system (W1B-HTK)

This system comprises two different subsystems, as depicted in Fig. 7: The ASR subsystem consists of an LVCSR system that produces a 1-best word sequence for each speech file. The STD subsystem consists of a word-based term detector and a decision maker that outputs reliable detections.

The ASR subsystem is built from the HTK tool [83]. First, the VAD of the Voicebox [130] tool is applied to segment the speech signal into speech segments. These speech segments are next decoded by the HTK tool to produce a 1-best word sequence as output. The ASR subsystem employs 39-dimensional PLP coefficients as acoustic features and three-state cross-word triphone models as acoustic models. The acoustic models have been trained from the Spanish partition of the EPPS corpus, which amounts to about 99 h of speech [131], and all the training/development data provided by the organizers except the MAVIR-02, MAVIR-07, and MAVIR-09 speech files. MAVIR-07 speech file was used as development file for parameter tuning, and the two other speech files were removed from the acoustic model training material since the STD performance on that development file degrades when both were used for acoustic model training. In total, about 101.5 h of speech were employed for acoustic model training. As a language model, a word trigram LM has

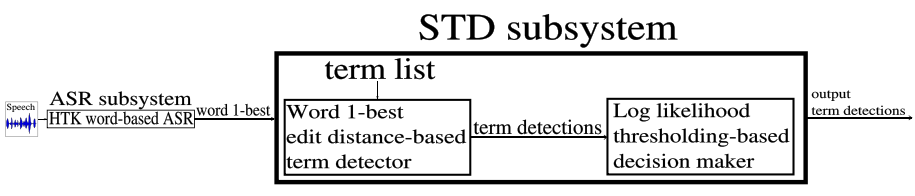

Fig. 7 Architecture of the W1B-HTK system

been employed. The LM was trained with the SRI-LM toolkit [127] from different text sources: (1) the Spanish Parliament partition (PARL) of the EPPS corpus used to train the acoustic models, which amounts to 17.5k words, (2) the training/development data provided by the organizers which amount to 5k words, and (3) data corresponding to different web pages whose topic relates to that of the MAVIR corpus (sentiment analysis, data crawling, etc.), and through web pages related to companies mentioned in the training/development data (daedalus, bitext, isoco, etc.), which amount to 7.2k words. In total, a vocabulary of 30k words has been used for LM training.

The STD subsystem comprises an edit distance-based term detector, which treats the term search in one way or another depending on the number of words the term consists of, and a decision maker that ascertains reliable detections. For term detection, an exact match in the 1-best word sequence is conducted in case the term is composed of a single or two words. For a term composed of three or more words, this is detected in case two of its words appear in the 1-best word sequence. For single-word and double-word terms, the start and end times of each detection are assigned the initial time of the first word and the end time of the last word, respectively. For terms with more than two words, the start and end times of the detection consider all the words, even those that are wrongly recognized. The confidence score for each detection is the sum of the scores given by the HTK tool to all the words of the term that are correctly recognized, divided by the number of words that are correctly recognized.

This system does not integrate a method to handle OOV terms, and hence terms absent from the HTK-based speech recognizer vocabulary cannot be detected.

For parameter tuning, this system employed the MAVIR-07 speech file from the training/development dataset. To do so, the entire STD system was first built and next applied on this file to obtain the optimal parameter set of the ASR and STD subsystems. Next, the optimal parameter set is applied on the test data to hypothesize the detections of the test term list.

Word lattice-based Kaldi ATWV-based STD system (WL-ATWV-Kaldi)

This system comprises two different subsystems, as depicted in Fig. 8: The ASR subsystem employs a Kaldi-based speech recognizer [68] to decode speech utterances and produce word lattices. The STD subsystem consists of the Kaldi term detector [68–70] to search for the input terms within the word lattices and the Kaldi decision maker [42] to output reliable detections.

For the ASR subsystem, a word-based speech recognizer using Kaldi toolkit [68] has been constructed. First, an energy-based VAD implemented in SoX has been employed to remove non-speech segments. For word-based speech recognition, 13-dimensional MFCCs with cepstral mean and variance normalization (CMVN) applied were used as acoustic features. The normalized MFCC features then pass a splicer which augments each frame by its left and right four neighboring frames. A linear discriminant analysis is then employed to reduce the feature dimension to 40, and a maximum likelihood linear transform (MLLT) is applied to match the diagonal assumption in GMM. These acoustic modeling training stages were chosen to maximize ATWV performance on training/development data. As acoustic models, context-dependent phone GMM/HMMs were trained with the training part of the Fisher Spanish corpus (about 268 h of conversational telephone speech) and the training/development data provided by the organizers (about 5 h of speech). For English words that appear in the Fisher Spanish corpus, a letter-to-sound module has been used to build the phone transcription of these words using Spanish phones. In total, about 273 h of speech have been used for acoustic model training. These acoustic models were augmented with some non-speech events present in the Fisher Spanish corpus-like background noise, laugh, breath, cough, sneeze, and lip-smack events, which were modeled as context-independent acoustic models. As a language modeling, a word trigram trained with a vocabulary of 30k words has been employed. This language model has been trained from the same data used to train the acoustic models.

The STD subsystem integrates the Kaldi term detector [68–70], as described in the fusion system, and the Kaldi decision maker [42]. This decision maker conducts the YES/NO decision for each detection based on the confidence score computed according to Eq. 2:

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

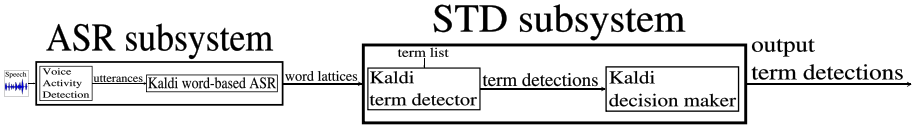

Fig. 8 Architecture of the WL-ATWV-Kaldi system

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

This system does not incorporate any mechanism to deal with the OOV terms.

This system employed all the training/development data for ASR subsystem training. Therefore, the parameter tuning was carried out as follows: First, the ASR subsystem was trained (acoustic and language models) with the Fisher Spanish corpus and five training/development files of the training/development data (all except MAVIR-02 and MAVIR-03). These two files (MAVIR-02 and MAVIR-03) were used for parameter tuning. Therefore, the entire STD system was run on MAVIR-02 and MAVIR-03 speech files and the optimal parameter set was obtained, including the type of acoustic models to use, which was chosen to maximize ATWV performance. This parameter set, along with the acoustic models, were finally used to hypothesize detections of the test term list. It must be noted that, as explained before, all the training/development data were used for acoustic and language model training in the ASR subsystem of the final system submitted. This aimed at building a more robust set of models for the evaluation.

Word lattice-based Kaldi WER-based STD system (WL-WER-Kaldi)

This system is the same as the WL-ATWV-Kaldi, with the only difference that acoustic models were optimized for word error rate (WER) performance on Fisher Spanish data. Acoustic modeling also includes maximum likelihood linear regression (MLLR) and speaker adaptive training (SAT) to improve model robustness. In addition, a discriminative training approach based on the maximum mutual information (MMI) criterion was employed to produce the final acoustic models used in the ASR subsystem.

Phone 1-best-based HTK STD system (P1B-HTK)

This system comprises two different subsystems, as depicted in Fig. 9: The ASR subsystem consists of a phone-based speech recognizer that decodes the speech utterances to generate phone lattices and 1-best phone sequences. The STD subsystem consists of a phone 1-best-based term detector and a fusion-based decision maker to output reliable detections.

The ASR subsystem is an HTK-based phone recognition system [83] (and therefore, not a word-based ASR system as the rest of the systems described before), which produces a phone lattice and a 1-best phone sequence for each speech file. The ASR subsystem employs 39-dimensional MFCCs as acoustic features with cepstral mean compensation and histogram equalization applied. Three-state context-dependent phone models have been trained from different speech sources: (1) the noise-free phonetically-balanced data of the ALBAYZIN database [132], which amount to 12.8 h of speech, (2) the close-talk microphone speech data from Speech-Dat-Car database [133], which amount to 18.85 h of speech, (3) data recorded with the close talk microphone and one of the lapel microphones of the Domolab database [134], which amount to 9.33 h of speech, and (4) data corresponding to the Spanish parliament sessions of the TC-STAR database [135], which amount to 111.89 h of speech. In total, about 153 h of speech have been used for acoustic model training. The text transcriptions of these speech sources have been used to train the phone trigram LM with the SRI-LM toolkit [127]. In total, there are about 500k words in the text material and about 3 million phones that are finally used to train the phone trigram.

The STD subsystem employs a term detector to output putative detections from the 1-best phone sequence and a fusion-based decision maker to ascertain reliable detections. For term detection, the 1-best phone sequence is used as source text for an edit distance search. In the search, each detection could be any substring which has a phone edit distance with the search term of less than 50 % of its length. Start and end times of each detection are assigned the start time of the first phone that is correctly recognized in the phone substring and the end time of the last phone that is correctly recognized in the phone substring. Detections that overlap in time are removed by keeping the best one (i.e., the one with the minimum edit distance). Once all the detections have been obtained, two different scores are assigned to each: one derived from the edit distance computed during term search, and an acoustic confidence measure obtained from the lattice. The former is computed from standard substitution, insertion,

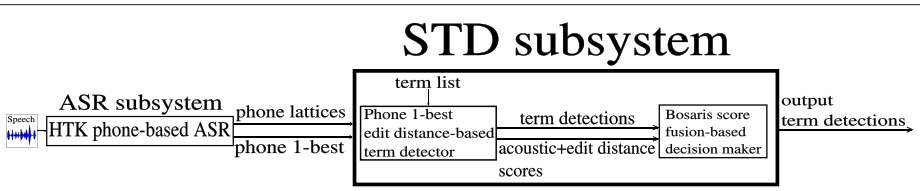

Fig. 9 Architecture of the P1B-HTK system

and deletion errors in the 1-best phone sequence and normalized by the length of the term. To obtain the acoustic confidence measure, the following steps are conducted: (1) the lattice is determinized with HTK tool [83], (2) an acoustic mesh graph of the phone lattice is obtained using the lattice tool of the SRI-LM toolkit [127], and (3) the confidence calculated in the acoustic mesh graph is used in a modified edit distance algorithm where, instead of all costs equal to 1, the confidence of the matching phones (those that are correctly recognized in the phone lattice) with the search term are accumulated. Then, the score of a detection is the sum of the confidences of the matching phones through the acoustic mesh of the search term between the time limits where the detection resides. This score is also normalized by the length of the term. The decision maker fuses these two scores with the Bosaris toolkit, produces the final confidence score for each detection, and ascertains reliable detections.

The entire set of training/development data was employed for parameter tuning. Therefore, the whole STD system was first built, and next this was run on training/development data to obtain the optimal parameter set. This optimal parameter set was next applied to hypothesize the detections corresponding to the test data.

It must be noted that this system is based on phone ASR and hence allowing for fast search and OOV term detection.

Results and discussion

System results are presented in Tables 10 and 11 for training/development and test data respectively. They show a different behavior, in terms of result ranking. The best performance on training/development data is obtained by the WL-WER-Kaldi system, whereas the best performance on test data is obtained by the WL-Kaldi system. Paired t-tests show that the best performance of the WL-WER-Kaldi system on training/development data is statistically significant compared to the rest of the systems ($p < 10^{-10}$) for ATWV. This better performance is due to the use of the training/development data for acoustic and language model training, which causes this system to be clearly biased towards these data. This is confirmed by the STD performance on training/development data obtained by the WL-ATWV-Kaldi system, which just differs from the WL-WER-Kaldi system in a few acoustic model training techniques. However, the WL-Kaldi system does not employ the training/development data for acoustic model training, and the language model does not include the decoded sentence, hence resulting in an unbiased system towards the training/development data. On the other hand, the P1B-HTK system, which employs a phone recognition-based STD system, obtains the worst performance on both sets of data, due to the absence of lexical information in the ASR subsystem. The rest of the systems significantly outperform the performance of this P1B-HTK system for training/development data ($p < 10^{-14}$) and test data ($p < 10^{-4}$) for ATWV.

WL-ATWV-Kaldi and WL-WER-Kaldi systems obtain different result ranking on both sets of data. The latter outperforms the former on training/development data and the contrary occurs on test data. Paired t-tests show that the better performance of the WL-WER-Kaldi system over the WL-ATWV-Kaldi system on training/development data is statistically significant ($p < 10^{-12}$) for ATWV, and the better performance of the WL-ATWV-Kaldi system over the WL-WER-Kaldi system is also statistically significant ($p < 10^{-5}$) for test data for ATWV. This is again, partially, due to the bias introduced in these systems with the use of the training/development data for acoustic and language model training. Although the WL-WER-Kaldi system employs more robust ASR techniques than the WL-ATWV-Kaldi system, these techniques only improved the WER of the ASR subsystem and obtained a worse STD performance on a subset of the training/development data provided by the organizers used for STD parameter tuning. This is also causing the worse performance on test data compared with the WL-ATWV-Kaldi system.

This difference in result ranking is also observed for W1B-HTK and WL-ATWV-Kaldi systems. On training/development data, the WL-ATWV-Kaldi system is biased towards these data (in terms of acoustic model and language model training), and hence better performance is obtained. This improvement on training/development data is statistically significant ($p < 10^{-9}$) for ATWV for a paired t-test. However, the W1B-HTK system, which

Table 10 Results of the STD ALBAYZIN 2014 evaluation on training/development data

System ID	MTWV	ATWV	p(FA)	p(Miss)	Precision	Recall	*F*-measure
Fusion	0.5676	0.5676	0.00007	0.363	0.8704	0.7385	0.7991
WL-Kaldi	0.5816	0.5816	0.00008	0.341	0.8604	0.7543	0.8039
W1B-HTK	0.4634	0.4622	0.00006	0.472	0.8483	0.5854	0.6927
WL-ATWV-Kaldi	0.6287	0.6233	0.00004	0.331	0.9391	0.6990	0.8014
WL-WER-Kaldi	0.8327	0.8155	0.00002	0.144	0.9773	0.8519	0.9103
P1B-HTK	0.0746	0.0746	0.00003	0.893	0.7093	0.1269	0.2153

Table 11 Results of the STD ALBAYZIN 2014 evaluation on test data

System ID	MTWV	ATWV	p(FA)	p(Miss)	Precision	Recall	*F*-measure
Fusion	0.4872	0.4868	0.00003	0.483	0.9666	0.6207	0.7560
WL-Kaldi	0.5451	0.5350	0.00008	0.374	0.9104	0.7473	0.8209
W1B-HTK	0.2026	0.1980	0.00016	0.642	0.7981	0.4927	0.6093
WL-ATWV-Kaldi	0.2018	0.1972	0.00007	0.731	0.8829	0.3194	0.4691
WL-WER-Kaldi	0.1389	0.1316	0.00003	0.828	0.8498	0.2561	0.3936
P1B-HTK	0.0391	0.0297	0.00004	0.917	0.7275	0.1222	0.2093

employs more sources that augment the variability in language model training and less training/development data for acoustic model training, results in a less biased system towards the training/development data. This causes similar performance for W1B-HTK and WL-ATWV-Kaldi systems on test data at the ATWV operating point. Moreover, for test data, the performance gap between both systems is not statistically significant ($p \approx 0.9$) for ATWV.

WL-Kaldi system performs the best on test data. This is due to two reasons: (1) it has the most robust ASR subsystem of those presented by the participants in terms of ASR techniques (SGMM for acoustic modeling), the type of the speech data (spontaneous speech) used for acoustic model training is very similar to that of the evaluation files, and the text material used for language model training comes from a large variety of text sources. (2) All the test terms were included within the ASR subsystem vocabulary, hence getting rid of the OOV term issue. Table 12 shows the OOV rate of the systems (i.e., the number of test terms that do not appear in the ASR subsystem vocabulary), which shows that the OOV rate plays an important role in system performance. Paired t-tests show that the best performance of the WL-Kaldi system is statistically significant compared to the rest of the systems ($p < 10^{-9}$) except the fusion system for ATWV. The fusion of this WL-Kaldi system with the other system presented in the fusion system yields the worse STD performance than the best system in isolation (WL-Kaldi system). This performance gap is statistically significant for a paired t-test ($p < 10^{-2}$) for ATWV. This suggests that a better fusion strategy is necessary.

Table 12 OOV rate of the word-based systems on test data. The P1B-HTK system is not presented in this table since it employs phone-based ASR

System ID	OOV rate (%)
Fusion	0 %
WL-Kaldi	0 %
W1B-HTK	12.87 %
WL-ATWV-Kaldi	13.86 %
WL-WER-Kaldi	13.86 %

It can also be observed in Table 10 that some systems show a slight degradation on training/development data when MTWV and ATWV are compared. This is because of the optimal threshold in the decision maker that was calculated from a subset of the training/development data and next applied to the whole set.

In terms of the F-measure values presented in Tables 10 and 11, similar trends are observed. The systems that obtained the best performance for ATWV metric are also the best for F-measure metric, with similar significance levels. The only difference in terms of result ranking relates to WL-Kaldi and WL-ATWV-Kaldi systems on training/development data. Whereas for ATWV the WL-ATWV-Kaldi system outperforms the WL-Kaldi system, the contrary occurs for F-measure. The improvement obtained by the WL-ATWV-Kaldi system for ATWV is statistically significant ($p < 10^{-2}$), and the improvement obtained by the WL-Kaldi system for F-measure is not statistically significant ($p \approx 0.7$). This difference in terms of system ranking is due to the different weight given to precision and recall in the ATWV and F-measure metrics. In addition, the improvement obtained by the W1B-HTK system over the WL-ATWV-Kaldi system on test data is statistically significant ($p < 10^{-4}$) for F-measure. Again, the different behavior of both metrics (ATWV and F-measure) is causing this discrepancy.

DET curves are presented in Figs. 10 and 11 for training/development data and test data, respectively. They show the system performance working at different miss/FA ratios. On test data, fusion and WL-Kaldi systems clearly outperform the rest of the STD systems for almost every operating point, as expected from the ATWV results. The fusion system outperforms the WL-Kaldi system for low FA rates and the contrary occurs for low miss rates. This means that the fusion system is still giving some benefit for low miss rates. The P1B-HTK system performs the worst for almost every operating point, except for low FA rates, where the W1B-HTK system (based on an HTK recognizer and a 1-best word search) performs the worst. When comparing the DET curves based on the ASR output (i.e., the term detection input), it can be seen that STD systems that employ word lattices (fusion, WL-Kaldi, WL-ATWV-Kaldi, and

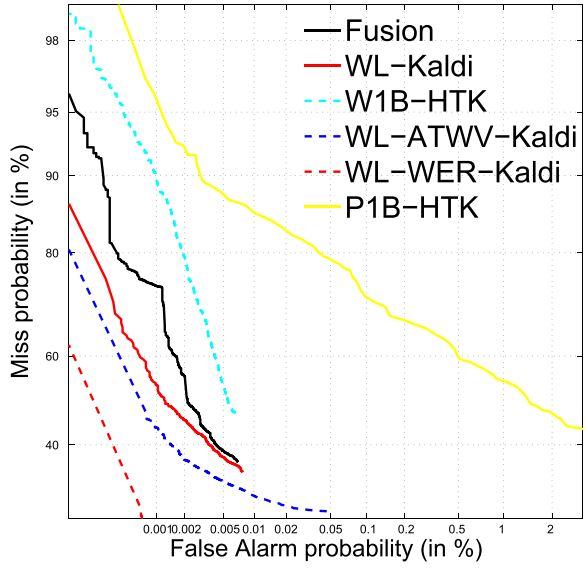

Fig. 10 DET curves of the STD systems for training/development data

WL-WER-Kaldi) outperform the 1-best word search of the W1B-HTK system, and the latter outperforms the phone recognition-based STD of the P1B-HTK system for most of the region. This is expected, since word lattices are able to keep more hypotheses for term detection than the 1-best word/phone sequences, despite of the better ATWV obtained by the W1B-HTK system over WL-ATWV-Kaldi and WL-WER-Kaldi systems. On training/development data, the bias observed in the ATWV results for WL-ATWV-Kaldi and WL-WER-Kaldi systems is clearly causing their best performance for every operating point. The P1B-HTK system, as in test data, also performs the worst on these data.

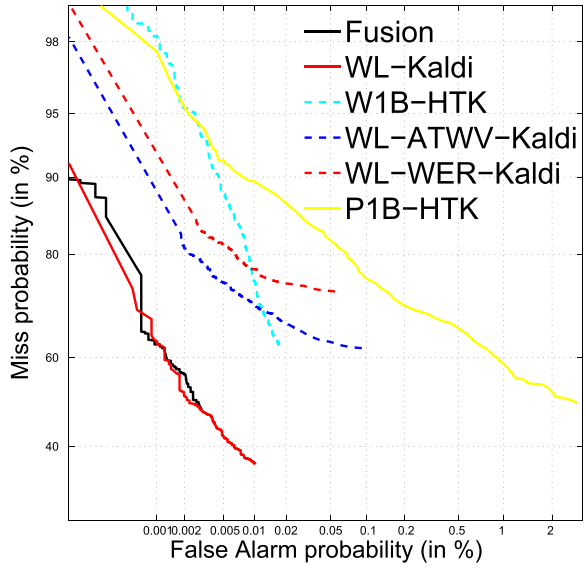

Fig. 11 DET curves of the STD systems for test data

Comparison to previous STD evaluations

Although our evaluation results cannot be directly compared to those obtained in previous NIST and NTCIR STD evaluations because the database/language/metric set used in our case is different, we can shed some light with respect to performance comparison to other STD evaluations held across the world. On the one hand, we can mention that our results are better than those reported for Arabic and Mandarin languages on the NIST STD 2006 evaluation (see Table 6). One possible reason for these results is that Spanish could be an *easier* language than Arabic and Mandarin from an ASR perspective (Spanish has very regular grapheme-to-phoneme mapping [120], Mandarin is a tonal language [136], which adds more complexity to the ASR system, diacritization in Arabic [137] adds more complexity to Arabic ASR systems). Other clearer reason is that the performance of the best system presented in this paper corresponds to a word-based system with no OOV terms. For English language and probably *easier* domains compared to that of the Spanish MAVIR corpus (telephone speech and broadcast news vs. oral talks in real workshops), the performance of those systems is better than that obtained in this evaluation. However, when the domain difficulty increases (roundtable meeting rooms for English) and all the terms are INV, system performance for Spanish on spontaneous speech (MAVIR corpus) is better than for English on meeting domain. Certainly, this is not the common scenario, and typically, OOV terms considerably degrade system performance. The systems presented here that have OOV terms are obtaining worse results than those obtained in the English roundtable meeting domain and the Arabic (except for broadcast news) and Mandarin languages in the NIST STD 2006 evaluation. This is mainly due to the difficulty inherent to the acoustic database conditions and the list of terms (containing foreign terms) employed in this evaluation.

On the other hand, the best systems submitted to the different NTCIR STD and Open KWS evaluations show, in general, similar performance rates (in terms of F-measure and ATWV metrics) to those obtained in our Spanish STD evaluation (see Tables 7 and 11 and "Comparison to other evaluations" section).

All these findings suggest that STD systems in highly difficult domains (e.g., real workshops) for Spanish language are, at least, as effective as for other languages/domains.

Performance analysis of STD systems based on term length

An analysis of the performance of the STD systems based on the length (in number of phonemes) of the test terms has been conducted and results are shown in Table 13.

Table 13 Results of the STD ALBAYZIN 2014 evaluation on test data based on the term length. "SHORT" denotes short-length terms, "MEDIUM" denotes medium-length terms, and "LONG" denotes long-length terms. The term length considers the number of phonemes of the given term

System ID	SHORT (< 8 phonemes)		MEDIUM (8–10 phonemes)		LONG (> 10 phonemes)	
	MTWV	ATWV	MTWV	ATWV	MTWV	ATWV
Fusion	0.4194	0.4194	0.5423	0.5415	0.4648	0.4648
WL-Kaldi	0.4929	0.4805	0.5841	0.5595	0.5576	0.5569
W1B-HTK	0.1440	0.1417	0.1794	0.1607	0.3457	0.3457
WL-ATWV-Kaldi	0.1685	0.1532	0.2250	0.2192	0.2256	0.2103
WL-WER-Kaldi	0.1540	0.1490	0.1345	0.1237	0.1521	0.1248
P1B-HTK	0.0638	0.0435	0.0375	0.0265	0.0199	0.0181

Test terms have been divided into three categories: short-length terms (terms with up to 7 phonemes), medium-length terms (terms between 8 and 10 phonemes), and long-length terms (terms with more than 10 phonemes). In general, longer terms should exhibit better performance than shorter terms, since these are naturally more confusable within speech data. However, this is not always the case.

For the word-based STD systems, it is clear from Table 13 that the medium-length term performance is better than that of the short-length terms. However, the STD performance degrades, in general, for long-length terms compared to medium-length terms. It must be noted that most of the long-length terms involve multi-word terms, which are more difficult to detect, in general, in STD (these can contain some OOV words, or some of their words are wrongly recognized in the ASR subsystem). A more detailed analysis for medium- and long-length terms composed of single- and multi-word terms is shown in Table 14. It is clear from Table 14 that the worse overall performance obtained by the long-length terms compared to medium-length terms is caused by those that contain two or more words. When comparing the medium- and long-length terms that contain only one word, the performance for long-length terms is better than for medium-length terms, which is the trend in STD for single-word terms.

From Table 13, the WL-Kaldi system performs the best for short-, medium-, and long-length terms, as expected from the overall STD results. Paired t-tests show that the improvement of this system over the rest is, in general, statistically significant for short-length terms ($p < 10^{-2}$ over the fusion system and $p < 10^{-6}$ over the rest), for medium-length terms ($p < 10^{-6}$ over all the systems except the fusion system), and for long-length terms ($p < 10^{-4}$ over the fusion system, $p < 10^{-3}$ over the W1B-HTK system, and $p < 10^{-5}$ over the rest).

When comparing the results of Table 14 across the systems, it is shown that, in general, the WL-Kaldi system also performs the best. Paired t-tests show that the improvement of this system over the rest is statistically significant ($p < 10^{-6}$ for all the systems except the fusion system for medium-length single-word terms, and $p < 10^{-4}$ for all the systems for long-length single-word terms). However, for terms involving multiple words, paired t-tests do not show any statistical difference between systems.

For the phone-based STD system (P1B-HTK), the performance degrades when the length of the term increases.

Table 14 Results of the STD ALBAYZIN 2014 evaluation on test data for single-word medium- and long-length terms and multi-word medium- and long-length terms. "MEDIUM-SINGLE" denotes medium-length terms that are composed of a single word, "LONG-SINGLE" denotes long-length terms that are composed of a single word, "MEDIUM-MULTI" denotes medium-length terms that are composed of two or more single words, and "LONG-MULTI" denotes long-length terms that are composed of two or more single words

System ID	MEDIUM-SINGLE		LONG-SINGLE		MEDIUM-MULTI		LONG-MULTI	
	MTWV	ATWV	MTWV	ATWV	MTWV	ATWV	MTWV	ATWV
Fusion	0.5664	0.5655	0.5813	0.5813	0	0	0.1250	0.1250
WL-Kaldi	0.6081	0.5789	0.7060	0.7050	0.1250	0.1250	0.1250	0.1250
W1B-HTK	0.1873	0.1679	0.4214	0.4214	0	0	0.1250	0.1250
WL-ATWV-Kaldi	0.2350	0.2290	0.3029	0.2824	0	0	0	0
WL-WER-Kaldi	0.1405	0.1292	0.2043	0.1676	0	0	0	0
P1B-HTK	0.0392	0.0277	0.0124	0.0100	0	0	0.0417	0.0417

First of all, it must be noted that for phone-based STD systems, there is no difference between single- and multi-word terms, since the detection subsystem relies on a phone sequence. The edit distance method employed in the P1B-HTK system for term detection causes that longer terms are more difficult to keep half of their phones in the 1-best phone sequence. Therefore, longer terms were more difficult to detect than shorter terms, despite these could generate more FAs.

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

A similar analysis based on the number of words of the test terms has been conducted and results are shown in Table 15. Similar conclusions to those mentioned in the previous section regarding single-word and multi-word terms are obtained. In general, for STD experiments on word-based systems, multi-word terms produce more errors in term detection. On the one hand, the OOV word problem affects in a greater extent multi-word terms. On the other hand, the word confusability inherent to ASR systems plays a more important role for multi-word terms, since these are composed of more than one word.

WL-Kaldi system performs the best both for single-word and multi-word term detection, as expected from the overall STD results. For single-word term detection, paired t-tests show that the improvement of this system over the rest is statistically significant ($p < 10^{-9}$ for all the systems except the fusion system, and $p < 10^{-4}$ for this). However, for multi-word term detection, paired t-tests show that the improvement of this system is only statistically significant compared with WL-ATWV-Kaldi and WL-WER-Kaldi systems ($p < 10^{-2}$). This shows the difficulty of multi-word term detection, even for word-based STD systems.

This analysis looses part of its meaning for phone-based STD systems such as P1B-HTK, since for these systems words do not exist, and hence single- and multi-word terms are just phone sequences.

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

An analysis of the performance of the STD systems based on in-vocabulary/out-of-vocabulary terms has been conducted and results are shown in Table 16. Typically, OOV terms are those absent from the ASR system vocabulary. However, since participants were allowed to use additional resources to train their systems, these may include OOV terms in the ASR component, in case these terms are learned from other sources. In case the traditional definition for OOV terms is applied in our case, systems would have different OOV terms, and hence a fair comparison would not be possible. In our case, in-vocabulary terms are those that appear in the training/development speech data provided by the organizers, and out-of-vocabulary terms are those that do not. It is clear from Table 16 that the system performance degrades for terms that do not appear in the training/development speech data. The only exception is the P1B-HTK system, which employs a phone-based ASR subsystem, and hence both in-vocabulary and out-of-vocabulary terms are treated equally (since the training/development data are provided by the organizers, and thus the in-vocabulary terms have not been used for acoustic model and phone-based language model training).

WL-Kaldi system achieves the best overall STD performance. Moreover, this system also obtains the best performance for in-vocabulary terms. Paired t-tests show that this improvement is statistically significant ($p < 10^{-8}$ for all but fusion system, and $p < 10^{-2}$ for fusion system). Therefore, it is clear that having the more robust ASR subsystem also plays a very important role in the overall performance, since all the in-vocabulary terms also appear in the ASR vocabulary of the rest of the systems.

It can be seen that the WL-ATWV-Kaldi system outperforms the W1B-HTK system for in-vocabulary terms and the contrary occurs for out-of-vocabulary terms. However, the improvement of the WL-ATWV-Kaldi system over the W1B-HTK system on in-vocabulary terms is not statistically significant for a paired t-test ($p \approx 0.5$),

Table 15 Results of the STD ALBAYZIN 2014 evaluation on test data for single-word terms and multi-word terms. "SINGLE" denotes single-word terms (i.e., terms that are composed of a single word) and "MULTI" denotes multi-word terms (i.e., terms that are composed of two or more single words)

System ID	SINGLE				MULTI			
	MTWV	ATWV	p(FA)	p(Miss)	MTWV	ATWV	p(FA)	p(Miss)
Fusion	0.5184	0.5180	0.00003	0.450	0.1471	0.1471	0	0.853
WL-Kaldi	0.5790	0.5680	0.00009	0.333	0.1765	0.1765	0	0.824
W1B-HTK	0.2132	0.2081	0.00017	0.617	0.0882	0.0882	0	0.912
WL-ATWV-Kaldi	0.2203	0.2153	0.00007	0.707	0	0	0	1
WL-WER-Kaldi	0.1516	0.1437	0.00004	0.812	0	0	0	1
P1B-HTK	0.0346	0.0243	0.00005	0.917	0.0882	0.0882	0	0.912

Table 16 Results of the STD ALBAYZIN 2014 evaluation on test data for in-vocabulary terms and out-of-vocabulary terms. "INV" denotes in-vocabulary terms and 'OOV' denotes out-of-vocabulary terms

System ID	INV				OOV			
	MTWV	ATWV	p(FA)	p(Miss)	MTWV	ATWV	p(FA)	p(Miss)
Fusion	0.5641	0.5636	0.00004	0.398	0.2652	0.2652	0.00001	0.730
WL-Kaldi	0.6189	0.6060	0.00010	0.282	0.3357	0.3304	0.00001	0.659
W1B-HTK	0.2243	0.2136	0.00020	0.572	0.1530	0.1530	0.00002	0.831
WL-ATWV-Kaldi	0.2518	0.2463	0.00008	0.668	0.0652	0.0556	0.00002	0.916
WL-WER-Kaldi	0.1800	0.1690	0.00009	0.732	0.0322	0.0237	0.00003	0.939
P1B-HTK	0.0366	0.0248	0.00005	0.916	0.0477	0.0436	0.00004	0.915

so both systems can be said to yield similar performance for in-vocabulary term detection. On the other hand, the improvement of the W1B-HTK system over the WL-ATWV-Kaldi system for OOV terms is statistically significant for a paired t-test ($p < 10^{-2}$). Given that in-vocabulary terms appear in the ASR subsystem vocabulary, both systems can detect them. However, the lower OOV rate of the W1B-HTK system compared to the WL-ATWV-Kaldi system (see Table 12) causes out-of-vocabulary term detection degrade in a greater extent in the WL-ATWV-Kaldi system. In addition, the bias of the WL-ATWV-Kaldi system towards training/development data can be enhancing the in-vocabulary term detection. This is confirmed by the degradation of the WL-WER-Kaldi system from in-vocabulary terms to out-of-vocabulary terms on test data, which is also caused by that bias.

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

An analysis of the performance of the STD systems in terms of Spanish (in-language) and foreign (out-of-language) test terms has been conducted and results are shown in Table 17. As expected, performance degradation is observed in foreign term detection. However, this degradation is not constant across the systems. fusion and WL-Kaldi systems yield the best performance for foreign terms, since these appear in the ASR subsystem vocabulary. Paired t-tests show that the WL-Kaldi system significantly improves ($p < 10^{-2}$) the rest of the systems except the fusion system ($p \approx 0.3$) for foreign term detection. The WL-ATWV-Kaldi system also maintains a relatively good performance for foreign terms compared to that obtained for in-language terms. All these foreign terms are in English, and this system used the English words that appear in the Fisher Spanish corpus for system training. This is clearly giving some benefit for foreign term detection. The WL-WER-Kaldi system has been trained with the same data. However, its worse overall performance compared to the WL-ATWV-Kaldi system is also producing worse performance in foreign term detection.

For in-language terms, the best performance is obtained by the WL-Kaldi system, as expected from the overall STD results. Paired t-tests show that the WL-Kaldi system significantly outperforms the rest of the systems for in-language term detection ($p < 10^{-9}$ for all the systems except the fusion system and $p < 10^{-3}$ for this).

Performance analysis of STD systems based on specific terms

Finally, an additional analysis of the performance of the STD systems for specific search terms has been conducted and results are shown in Tables 18 and 19. Table 18 shows the results of a randomly selected set of three terms with different length, being in-language, in-vocabulary, and

Table 17 Results of the STD ALBAYZIN 2014 evaluation on test data for in-language and out-of-language (foreign) terms. "IN-LANG" refers to Spanish terms and "OUT-LANG" refers to foreign terms

System ID	IN-LANG				OUT-LANG			
	MTWV	ATWV	p(FA)	p(Miss)	MTWV	ATWV	p(FA)	p(Miss)
Fusion	0.5030	0.5025	0.00003	0.466	0.2575	0.2575	0.00001	0.732
WL-Kaldi	0.5580	0.5475	0.00008	0.364	0.3651	0.3545	0.00003	0.603
W1B-HTK	0.2139	0.2097	0.00016	0.623	0.0393	0.0287	0.00004	0.918
WL-ATWV-Kaldi	0.2094	0.2045	0.00007	0.723	0.1339	0.0916	0.00001	0.855
WL-WER-Kaldi	0.1458	0.1382	0.00003	0.820	0.0389	0.0353	0.00002	0.940
P1B-HTK	0.0418	0.0317	0.00005	0.911	0	0	0	1

Table 18 Results of the STD ALBAYZIN 2014 evaluation on test data based on term length for specific search terms. "Term 1" is short-length, "Term 2" is medium-length, and "Term 3" is long-length. All are single-word, in-vocabulary, and in-language terms. 'OCC' stands for the OCC value for each given term computed with the NIST STD evaluation tool

System ID	Term 1 (short)			Term 2 (medium)			Term 3 (long)		
	OCC	Precision	Recall	OCC	Precision	Recall	OCC	Precision	Recall
Fusion	0.665	0.9730	0.6667	0.738	0.8571	0.7500	0.843	0.8571	0.8571
WL-Kaldi	0.735	0.9302	0.7407	0.850	0.7778	0.8750	0.986	0.8750	1
W1B-HTK	0.480	0.9630	0.4815	0.738	0.8571	0.7500	0.843	0.8571	0.8571
WL-ATWV-Kaldi	0.331	0.9474	0.3333	0.613	0.8333	0.6250	0.843	0.8571	0.8571
WL-WER-Kaldi	0.348	0.9048	0.3519	0.487	0.8000	0.5000	0.557	0.8000	0.5714
P1B-HTK	0.165	0.9000	0.1667	0.113	0.5000	0.1250	0.000	0	0

single-word terms. The results show that long single-word terms are much more easily detected with the word-based systems than with the phone-based system (P1B-HTK) submitted to the evaluation. On the contrary, for the P1B-HTK system, long terms are more difficult to detect due to the search on the 1-best phone sequence, which poses more difficulties in finding hits. This confirms the findings observed from the results presented in Tables 13 and 14. In the same way, longer terms exhibit better performance than shorter terms (medium- and short-length terms) for word-based systems, as observed from the results in Tables 13 and 14.

Table 19 shows the system behavior for some term properties, which are varied from one term to another. In this case, we concentrate on medium-length terms, and the in-language/out-of-language, single-word/multi-word, and in-vocabulary/out-of-vocabulary properties are varied one at each time. These results show that OOV, multi-word terms are the most difficult to detect, even for word-based systems, which confirms the findings observed in the results presented in Tables 15 and 16. On the other hand, the results obtained with the P1B-HTK system confirm the difficulty in retrieving medium-length terms for the phone-based system submitted to the evaluation when the term properties produce a more *difficult* term.

Lessons learned

The spoken term detection ALBAYZIN 2014 evaluation is integrated within a more general search on speech ALBAYZIN evaluation. This is the second edition of the search on speech ALBAYZIN evaluation, after that held in 2012. Since then, the evaluation has involved different tasks. In the first edition (held in 2012), we organized three different tasks, named query-by-example spoken term detection, keyword spotting, and spoken term detection. In the first evaluation, however, most participants only submitted systems to the query-by-example STD evaluation [37] and only one participant submitted a system for STD and keyword spotting tasks. In this second edition (held in 2014), there was an additional task, named query-by-example spoken document retrieval. However, none of the participants submitted any system to this task, being most of them involved in the STD task. There were also some systems submitted to the query-by-example STD and keyword spotting tasks. Therefore, this is the first time that we really have an evaluation with several teams working on STD in Spanish.

Table 19 Results of the STD ALBAYZIN 2014 evaluation on test data based on specific search terms. "Term 1" is a multi-word, out-of-vocabulary, and in-language term, "Term 2" is a single-word, in-vocabulary, and foreign term, and "Term 3" is a single-word, out-of-vocabulary, and in-language term. All are medium-length terms. "OCC" stands for the OCC value for each given term computed with the NIST STD evaluation tool [40]

System ID	Term 1			Term 2			Term 3		
	OCC	Precision	Recall	OCC	Precision	Recall	OCC	Precision	Recall
Fusion	0	0	0	0.167	1	0.1667	0.600	1	0.6000
WL-Kaldi	0.500	1	0.5000	0.500	1	0.5000	0.900	1	0.9000
W1B-HTK	0	0	0	0.317	0.6667	0.3333	0.900	1	0.9000
WL-ATWV-Kaldi	0	0	0	0.333	1	0.3333	0	0	0
WL-WER-Kaldi	0	0	0	0.167	1	0.1667	0	0	0
P1B-HTK	0	0	0	0	0	0	0	0	0

Besides the possibility of evaluating results, conducting system comparison, and having a common framework to foster research in search on speech for Spanish, the organization of the evaluation has provided us several lessons that will be very useful for the organization of future evaluations (either by us or by other prospective organizers).

First of all, from the number of participants and the number of tasks that had very few or no submissions, it is crucial for future evaluations either to have one compulsory task or even to focus on a single task to concentrate research efforts. It could be useful to have a previous survey (as it is usually done in MediaEval evaluations) to select one or two tasks. Secondly, taking part in this kind of evaluation implies a considerable amount of work that sometimes is not as fruitful as expected. We consider that it is important to lower the entrance barriers for taking part in these evaluations. In this sense, the new i-vector Challenges launched by NIST in speaker and language recognition in the last years are good examples. The i-vector Challenges transform the speaker and language recognition tasks in a pattern recognition task (only i-vectors and no speech are provided) for which specific knowledge of speech processing is not required. Besides, evaluation organizers also provide a baseline system with relatively good performance that can be used to test improvements over a basic algorithm. This has considerably increased the participation in these evaluations. In our evaluation, and in particular in the spoken term detection task, the highest barrier is probably the difficulty in having a good LVCSR system in Spanish. By providing the lattices generated by a reasonable good speech recognition system for training, development and test data (as done in NTCIR STD evaluations), more research groups (apart from those working on speech processing) would be able to participate. We could also provide baseline systems to help researchers to focus on improvements rather than on building a functional system from scratch, which in some cases was the main goal of the participants in the evaluation.

Regarding the data preparation, we have been able to use the database of MAVIR project consisting of recordings of seminars and roundtables organized at the general meetings of the project (at large conference rooms with about 100 people). This database has resulted very challenging with many interesting properties (i.e., different noise levels, different speakers, foreign words, etc.). For instance, in the first edition of the search on speech ALBAYZIN evaluation held in 2012, we focused on single-word terms in Spanish, but in the second edition, we added multi-word terms and foreign terms in order to analyze the influence of these in system performance. The database was transcribed and aligned at the utterance level. This was very helpful to produce the manual term alignments, but even using this information, it took a considerable amount of time to produce the manual alignments. Although MAVIR data have been very useful, we consider that it will be necessary to use additional data (for instance from broadcast news or perhaps more challenging TV programs) to make the evaluations evolve and not become repetitive. We are currently preparing more data in order to perform a new and more challenging evaluation in 2016. Besides using new data, we will probably reuse the same MAVIR data to assess technology improvements on a comparable basis.

In these two evaluation editions, we have prepared a training and development dataset and a test dataset. This has lead each participant to form the training data and development data in a different way, which has significantly complicated the comparison of the system results on this dataset, since the amount of data used for system training and system tuning is not consistent across participants. To solve this issue, three different datasets will be provided in future editions: training dataset, development dataset, and test dataset.

Performance of OOV terms is crucial in spoken term detection because OOV terms will inevitably occur when searching on speech. In this edition, we have conducted an analysis of the performance obtained by the different systems with respect to OOV terms. However, we have found significant differences in terms of OOV rate across the systems because we did not put any restrictions in the corpora that could be used for training the ASR component of the whole STD system. For future editions, it will be very helpful to define in advance and communicate to the participants the set of OOV terms that cannot be used for training in any way, so that OOV terms are actually OOV terms for all the systems.

The current STD evaluation focused on searching a training/development term list in training/development speech data and searching a test term list in test speech data. In future evaluations, the cross-data term search should be also considered. To do so, organizers should provide the alignments of the training/development terms in the test speech data and the alignments of the test terms in the training/development speech data. The purpose of this cross-data search is to see how critical tuning is for the different systems. For example, searching test terms in training/development speech data could be enhanced by unsupervised adaptation, whereas searching training/development terms in test speech data will measure the generalization capability of the systems on unseen data with the same term list for which good classifiers could have been developed.

Finally, in future editions, we would like to include other performance measures in the evaluation plan. In this evaluation, we only considered MTWV and ATWV. We have next included precision, recall, and the F-measure

in the analysis of the results, but this was not planned in advance and was not used for the evaluation itself. For future editions, we would also like to allow the participants to submit calibrated likelihood ratios as well as non-calibrated scores in order to measure calibration as well as other figures of merit such as the normalized cross entropy cost (Cnxe) employed in the last query-by-example search on speech task of MediaEval evaluation [27].

Preparing and running an evaluation is easier than taking part in it, as long as you have the proper data and funding, or even with very limited or no funding. Therefore, we hope to organize a new edition in 2016 but, in case it is impossible for us, at least we hope that these lessons could be useful for prospective organizers.

Conclusions

We have presented the spoken term detection ALBAYZIN 2014 evaluation, which is, to the best of our knowledge, the first STD evaluation in Spanish, as well as the six systems submitted. Four different research groups (GTM, GTH, ATVS-GEINTRA, and VivoLab) took part in the evaluation. There were two different types of systems submitted to the evaluation: word-based STD systems and a phone-based STD system. Five systems rely on a word-based speech recognizer and a subsequent search within the word lattice or 1-best word sequence of the evaluation terms. The other is based on a phone recognizer and a search in the 1-best phone sequence from an edit distance approach. Although the phone-based system performance is the worst, it allows for fast indexing and search. Given the challenging database chosen for the evaluation, results show a high performance for the best system (ATWV=0.5350) for which all the search test terms were included in the ASR system vocabulary. The other word-based STD systems, which suffer from the OOV word problem, exhibit serious performance degradation.

We have also shown that long single-word terms and in-vocabulary terms yield better STD performance. Contrary, multi-word terms and foreign terms tend to decrease the STD performance, as expected.

This is the first STD evaluation that has been conducted for Spanish language so far, which represents a good baseline for future research in this language. In addition, the database conditions (spontaneous speech and challenging audio conditions) chosen for the experiments and the highly heterogeneous list of terms (single- and multi-word terms, in-vocabulary and out-of-vocabulary terms, and in-language and foreign terms) make the evaluation and the database attractive enough for future research. Results presented in this paper indicate that there is still ample room for improvement when the list of terms contains a reasonable OOV rate (for 13 % OOV rate, the best performance obtained is ATWV=0.1980). The best result presented in this evaluation will be also considered as an interesting baseline for systems that do not have OOV terms. These results encourage us to maintain this evaluation in the next ALBAYZIN evaluation campaigns, trying to improve several issues that have arisen from the experience of this edition.

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

Authors' contributions
Systems submitted to the first spoken term detection evaluation for Spanish language are presented. Systems are categorized into word-based STD systems and phone-based STD systems. Analysis of system results is presented. Lessons learned from the STD evaluation are presented.

Acknowledgements
This work has been partly supported by project CMC-V2 (TEC2012-37585-C02-01) from the Spanish Ministry of Economy and Competitiveness. This research was also funded by the European Regional Development Fund, the Galician Regional Government (GRC2014/024, "Consolidation of Research Units: AtlantTIC Project" CN2012/160).

Author details
[1]GEINTRA, Universidad de Alcalá, Campus Universitario. Ctra. Madrid-Barcelona, km.33,600, Alcalá de Henares, Madrid, Spain. [2]Biometric Recognition Group - ATVS, Universidad Autónoma de Madrid, Av. Francisco Tomás y Valiente, 11. Escuela Politécnica Superior, Madrid, Spain. [3]Multimedia Technologies Group (GTM), AtlantTIC Research Center, E. E. Telecomunicación, Campus Universitario de Vigo, s/n, Vigo, Spain. [4]Speech Technology Group (GTH), Universidad Politécnica de Madrid, Depto. de Ingeniería Electrónica, ETSI de Telecomunicación, Madrid, Spain. [5]Voice Input Voice Output Laboratory (ViVoLab), Aragon Institute for Engineering Research (I3A), University of Zaragoza, C. María de Luna, 1. Ada Byron Building, Room GTCII, Zaragoza, Spain.

Received: 19 March 2015 Accepted: 6 July 2015
Published online: 07 August 2015

References

1. J Mamou, B Ramabhadran, O Siohan, in *Proc. of ACM SIGIR*. Vocabulary independent spoken term detection (ACM New York, NY, USA, 2007), pp. 615–622
2. J Mamou, B Ramabhadran, in *Proc. of Interspeech*. Phonetic query expansion for spoken document retrieval (ACM New York, NY, USA, 2008), pp. 2106–2109
3. D Can, E Cooper, A Sethy, C White, B Ramabhadran, M Saraclar, in *Proc. of ICASSP*. Effect of pronunciations on OOV queries in spoken term detection (Taipei, 2009), pp. 3957–3960
4. JG Fiscus, J Ajot, JS Garofolo, G Doddingtion, in *Proc. of Workshop on Searching Spontaneous Conversational Speech*. Results of the 2006 spoken term detection evaluation (ACM, New York, USA, 2007), pp. 45–50
5. D Vergyri, A Stolcke, RR Gadde, W Wang, in *Proc. of NIST Spoken Term Detection Workshop (STD 2006)*. The SRI 2006 spoken term detection system (NIST, Gaithersburg, MD, USA, 2006), pp. 1–15
6. D Vergyri, I Shafran, A Stolcke, RR Gadde, M Akbacak, B Roark, W Wang, in *Proc. of Interspeech*. The SRI/OGI 2006 spoken term detection system (ISCA, Baixas, France, 2007), pp. 2393–2396
7. M Akbacak, D Vergyri, A Stolcke, in *Proc. of ICASSP*. Open-vocabulary spoken term detection using graphone-based hybrid recognition systems (Las Vegas, NV, 2008), pp. 5240–5243
8. I Szöke, M Fapšo, M Karafiát, L Burget, F Grézl, P Schwarz, O Glembek, P Matějka, J Kopecký, J Černocký, in *Machine Learning for Multimodal Interaction*. Spoken term detection system based on combination of LVCSR and phonetic search, vol. 4892/2008, (2008), pp. 237–247. doi:10.1007/978-3-540-78155-4_21

9. I Szöke, L Burget, J Černocký, M Fapšo, in *Proc. of SLT*. Sub-word modeling of out of vocabulary words in spoken term detection (Goa, 2008), pp. 273–276
10. I Szöke, M Fapšo, L Burget, Černocký, in *Proc. of Speech Search Workshop at SIGIR*. Hybrid word-subword decoding for spoken term detection (ACM, New York, USA, 2008), pp. 42–48
11. S Meng, P Yu, J Liu, F Seide, in *Proc. of ICASSP*. Fusing multiple systems into a compact lattice index for Chinese spoken term detection (Las Vegas, NV, 2008), pp. 4345–4348
12. K Thambiratmann, S Sridharan, Rapid yet accurate speech indexing using dynamic match lattice spotting. IEEE Trans. Audio Speech Lang. Process. **15**(1), 346–357 (2007)
13. R Wallace, R Vogt, B Baker, S Sridharan, in *Proc. of ICASSP*. Optimising figure of merit for phonetic spoken term detection (Dallas, TX, 2010), pp. 5298–5301
14. C Parada, A Sethy, M Dredze, F Jelinek, in *Proc. of Interspeech*. A spoken term detection framework for recovering out-of-vocabulary words using the web (ISCA, Baixas, France, 2010), pp. 1269–1272
15. A Jansen, K Church, H Hermansky, in *Proc. of Interspeech*. Towards spoken term discovery at scale with zero resources (ISCA, Baixas, France, 2010), pp. 1676–1679
16. C Parada, A Sethy, B Ramabhadran, in *Proc. of ICASSP*. Balancing false alarms and hits in spoken term detection (Dallas, TX, 2010), pp. 5286–5289
17. D Schneider, T Mertens, M Larson, J Kohler, in *Proc. of Interspeech*. Contextual verification for open vocabulary spoken term detection, (2010), pp. 697–700. doi:10.1007/s11390-012-1228-x
18. C-A Chan, L-S Lee, in *Proc. of Interspeech*. Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping (Prague, 2010), pp. 693–696
19. C-P Chen, H-Y Lee, C-F Yeh, L-S Lee, in *Proc. of Interspeech*. Improved spoken term detection by feature space pseudo-relevance feedback (Prague, 2010), pp. 1672–1675
20. P Motlicek, F Valente, P Garner, in *Proc. of Interspeech*. English spoken term detection in multilingual recordings (Switzerland, 2010), pp. 206–209
21. NIST, The Ninth Text REtrieval Conference (TREC 9) (2000). http://trec.nist.gov
22. NIST, *The Spoken Term Detection (STD) 2006 Evaluation Plan*, 10th edn. (National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 2006). National Institute of Standards and Technology (NIST) . http://www.nist.gov/speech/tests/std
23. T Sakai, H Joho, in *Proceedings of NTCIR-9*. Overview of NTCIR-9 (National Institute of Informatics (NII), Tokyo, Japan, 2011), pp. 1–7
24. N Rajput, F Metze, in *Proceedings of MediaEval*. Spoken web search (CEUR Workshop Proceedings, Aachen, Germany, 2011), pp. 1–2
25. F Metze, E Barnard, M Davel, C van Heerden, X Anguera, G Gravier, N Rajput, in *Proceedings of MediaEval*. Spoken web search (CEUR Workshop Proceedings, Aachen, Germany, 2012), pp. 1–2
26. X Anguera, F Metze, A Buzo, I Szoke, LJ Rodriguez-Fuentes, in *Proceedings of MediaEval*. The spoken web search task (CEUR Workshop Proceedings, Aachen, Germany, 2013), pp. 1–2
27. X Anguera, LJ Rodriguez-Fuentes, I Szoke, A Buzo, F Metze, in *Proceedings of MediaEval*. Query by example search on speech at Mediaeval 2014 (Spain, 2014), pp. 1–2
28. H Joho, T Sakai, in *Proceedings of NTCIR-10*. Overview of NTCIR-10 (National Institute of Informatics (NII), Tokyo, Japan, 2013). pp. 1–7
29. H Joho, K Kishida, in *Proceedings of NTCIR-11*. Overview of the NTCIR-11 SpokenQuery&Doc task (National Institute of Informatics (NII), Tokyo, Japan, 2014), pp. 1–7
30. NIST, *OpenKWS13 Keyword Search Evaluation Plan*. (National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 2013). National Institute of Standards and Technology (NIST). http://www.nist.gov/itl/iad/mig/openkws13.cfm
31. NIST, *Draft KWS14 Keyword Search Evaluation Plan*. (National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 2013). National Institute of Standards and Technology (NIST). http://www.nist.gov/itl/iad/mig/upload/KWS14-evalplan-v11.pdf
32. Red Temática en Tecnologías del Habla. http://www.rthabla.es/
33. ISCA Special Interest Group: Iberian Languages (IL-SIG). http://www.isca-speech.org/iscaweb/index.php/sigs?layout=edit&id=132
34. B Taras, C Nadeu, Audio segmentation of broadcast news in the ALBAYZIN-2010 evaluation: overview, results, and discussion. EURASIP J. Audio Speech Music Process. **1**, 1–10 (2011)
35. M Zelenák, H Schulz, J Hernando, Speaker diarization of broadcast news in ALBAYZIN 2010 evaluation campaign. EURASIP J. Audio Speech Music Process. **19**, 1–9 (2012)
36. LJ Rodríguez-Fuentes, M Penagarikano, A Varona, M Díez, G Bordel, in *Proceedings of Interspeech*. The ALBAYZIN 2010 Language Recognition Evaluation (Waikoloa HI, 2011), pp. 1529–1532
37. J Tejedor, DT Toledano, X Anguera, A Varona, LF Hurtado, A Miguel, J Colás, Query-by-example spoken term detection ALBAYZIN 2012 evaluation: overview, systems, results, and discussion. EURASIP, J. Audio Speech Music Process. **23**, 1–17 (2013)
38. F Méndez, L Docío, M Arza, F Campillo, in *Proceedings of FALA*. The ALBAYZIN 2010 text-to-speech evaluation (Italy, 2010), pp. 317–340
39. IberSPEECH 2014 "VIII Jornadas en Tecnologías del Habla" and "IV Iberian SLTech Workshop". http://iberspeech2014.ulpgc.es/
40. NIST, *Evaluation Toolkit (STDEval) Software*. (National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 1996). National Institute of Standards and Technology (NIST). http://www.itl.nist.gov/iad/mig/tests/std/tools
41. I Szöke, M Fapšo, M Karafiát, L Burget, F Grézl, P Schwarz, O Glembek, P Matějka, S Kontár, J Černocký, in *Proc. of NIST Spoken Term Detection Evaluation Workshop (STD'06)*. BUT system for NIST STD 2006 - English (NIST, Gaithersburg, MD, USA, 2006), pp. 1–15
42. DRH Miller, M Kleber, C-L Kao, O Kimball, T Colthurst, SA Lowe, RM Schwartz, H Gish, in *Proc. of Interspeech*. Rapid and accurate spoken term detection (Belgium, 2007), pp. 314–317
43. H Li, J Han, T Zheng, G Zheng, in *Proc. of Interspeech*. A novel confidence measure based on context consistency for spoken term detection, (2012), pp. 2430–2433. http://doi.org/10.1587/transinf.E97.D.554
44. H-Y Lee, L-S Lee, Enhanced spoken term detection using support vector machines and weighted pseudo examples. IEEE Trans. Audio Speech Lang. Process. **21**(6), 1272–1284 (2013)
45. J Chiu, A Rudnicky, in *Proc. of Interspeech*. Using conversational word bursts in spoken term detection (ISCA, Baixas, France, 2013), pp. 2247–2251
46. F Seide, P Yu, C Ma, E Chang, in *Proc. of ICASSP*. Vocabulary-independent search in spontaneous speech (Montreal, 2004), pp. 253–256
47. B Logan, J-MV Thong, PJ Moreno, Approaches to reduce the effects of OOV queries on indexed spoken audio. IEEE Trans. Multimedia. **7**(5), 899–906 (2005)
48. B Logan, P Moreno, O Deshmuk, in *Proc. of HLT*. Word and sub-word indexing approaches for reducing the effects of OOV queries on spoken audio (San Francisco, CA, USA, 2002), pp. 31–35
49. B Ma, H Li, in *Proc. of ACM SIGIR*. A phonotactic-semantic paradigm for automatic spoken document classification (ACM, New York, USA, 2005), pp. 369–376
50. J Pinto, I Szöke, SRM Prasanna, H Heřmanský, in *Proc. of ACM SIGIR*. Fast approximate spoken term detection from sequence of phonemes (ACM, New York, USA, 2008), pp. 28–33
51. T Ohno, T Akiba, in *Proc. of Interspeech*. DTW-distance-ordered spoken term detection (ISCA, Baixas, France, 2013), pp. 3737–3741
52. S Nakagawa, K Iwami, Y Fujii, K Yamamoto, A robust/fast spoken term detection method based on a syllable n-gram index with a distance metric. Speech Commun. **55**(3), 470–485 (2013)
53. H Su, J Hieronymus, Y He, E Fosler-Lussier, S Wegmann, in *Proc. of SLT*. Syllable based keyword search: transducing syllable lattices to word lattices (South Lake Tahoe, NV, 2014), pp. 489–494
54. D Wang, J Frankel, J Tejedor, S King, in *Proc. of ICASSP*. A comparison of phone and grapheme-based spoken term detection (Las Vegas, NV, 2008), pp. 4969–4972
55. R Wallace, R Vogt, S Sridharan, in *Proc. of Interspeech*. A phonetic search approach to the 2006 NIST spoken term detection evaluation (Belgium, 2007), pp. 2385–2388
56. S Parlak, M Saraçlar, in *Proc. of ICASSP*. Spoken term detection for Turkish broadcast news (Las Vegas, NV, 2008), pp. 5244–5247
57. N Kanda, R Takeda, Y Obuchi, in *Proc. of SLT*. Using rhythmic features for japanese spoken term detection (Miami, FL, 2012), pp. 170–175

58. M Wollmer, B Schuller, G Rigoll, Keyword spotting exploiting long short-term memory. Speech Commun. **55**(2), 252–265 (2013)
59. J Tejedor, DT Toledano, D Wang, S King, J Colás, Feature analysis for discriminative confidence estimation in spoken term detection. Comput. Speech Lang. **28**(5), 1083–1114 (2014)
60. DA James, in *Proc. of ICASSP*. A system for unrestricted topic retrieval from radio news broadcasts (Atlanta, GA, 1994), pp. 279–282
61. GJF Jones, JT Foote, K Spärck Jones, SJ Young, in *Proc. of ACM SIGIR*. Retrieving spoken documents by combining multiple index sources (ACM, New York, USA, 1996), pp. 30–38
62. M Saraclar, R Sproat, in *Proc. of HLT-NAACL*. Lattice-based search for spoken utterance retrieval (Association for Computational Linguistics (ACL), Stroudsburg, PA, USA, 2004), pp. 129–136
63. K Iwata, K Shinoda, S Furui, in *Proc. of Interspeech*. Robust spoken term detection using combination of phone-based and word-based recognition (ISCA, Baixas, France, 2008), pp. 2195–2198
64. J Li, X Wang, B Xu, in *Proc. of Interspeech*. An empirical study of multilingual and low-resource spoken term detection using deep neural networks (ISCA, Baixas, France, 2014), pp. 1747–1751
65. P Yu, F Seide, in *Proc. of ICSLP*. A hybrid word/phoneme-based approach for improved vocabulary-independent search in spontaneous speech (ISCA, Baixas, France, 2004), pp. 293–296
66. A Yazgan, M Saraclar, in *Proc. of ICASSP*. Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition, (2004), pp. 745–748. doi:10.1109/ICASSP.2004.1326093
67. D Xu, F Metze, in *Proc. of Interspeech*. Word-based probabilistic phonetic retrieval for low-resource spoken term detection (Singapore, 2014), pp. 2774–2778
68. D Povey, A Ghoshal, G Boulianne, L Burget, O Glembek, N Goel, M Hannemann, P Motlicek, Y Qian, P Schwarz, J Silovsky, G Stemmer, K Vesely, in *Proc. of ASRU*. The KALDI speech recognition toolkit (IEEE, USA, 2011)
69. G Chen, S Khudanpur, D Povey, J Trmal, D Yarowsky, O Yilmaz, in *Proc. of ICASSP*. Quantifying the value of pronunciation lexicons for keyword search in low resource languages (Vancouver, BC, 2013), pp. 8560–8564
70. VT Pham, NF Chen, S Sivadas, H Xu, I-F Chen, C Ni, ES Chng, H Li, in *Proc. of SLT*. System and keyword dependent fusion for spoken term detection (South Lake Tahoe, NV, 2014), pp. 430–435
71. G Chen, O Yilmaz, J Trmal, D Povey, S Khudanpur, in *Proc. of ASRU*. Using proxies for OOV keywords in the keyword search task (Olomouc, 2013), pp. 416–421
72. IARPA, *Babel Program*. (Intelligence Advanced Research Projects Activity (IARPA), Washington DC, USA, 2011). Intelligence Advanced Research Projects Activity (IARPA). http://www.iarpa.gov/images/files/programs/babel/Babel-Kickoff-Summary.pdf
73. D Karakos, R Schwartz, in *Proc. of Interspeech*. Subword and phonetic search for detecting out-of-vocabulary keywords (ISCA Baixas, France, 2014), pp. 2469–2473
74. Y Wang, F Metze, in *Proc. of Interspeech*. An in-depth comparison of keyword specific thresholding and sum-to-one score normalization, (2014), pp. 2474–2478. doi:10.1007/978-3-319-11581-8_7
75. D Karakos, I Bulyko, R Schwartz, S Tsakalidis, L Nguyen, J Makhoul, in *Proc. of ICASSP*. Normalization of phonetic keyword search scores (Florence, 2014), pp. 7834–7838
76. WH Press, SA Teukolsky, T Vetterling, BP Flannery, *Numerical recipes: The art of scientific computing*. (Cambridge University Press, 2007)
77. J Chiu, A Rudnicky, in *Proc. of Interspeech*. Using conversational word bursts in spoken term detection (ISCA, Baixas, France, 2013), pp. 2247–2251
78. I-F Chen, NF Chen, C-H Lee, in *Proc. of Interspeech*. A keyword-boosted sMBR criterion to enhance keyword search performance in deep neural network based acoustic modeling (ISCA, Baixas, France, 2014), pp. 2779–2783
79. SP Rath, KM Knill, A Ragni, MJF Gales, in *Proc. of Interspeech*. Combining tandem and hybrid systems for improved speech recognition and keyword spotting on low resource languages (ISCA, Baixas, France, 2014), pp. 835–839
80. KM Knill, MJF Gales, SP Rath, PC Woodland, C Zhang, S-X Zhang, in *Proc. of ASRU*. Investigation of multilingual deep neural networks for spoken term detection (IEEE, USA, 2013), pp. 138–143
81. MJF Gales, KM Knill, A Ragni, SP Rath, in *Proc. of Spoken Language Technologies for Under-Resourced Languages*. Speech recognition and keyword spotting for low resource languages: Babel project research at CUED (ISCA, Baixas, France, 2014), pp. 16–23
82. KM Knill, MJF Gales, A Ragni, SP Rath, in *Proc. of Interspeech*. Language independent and unsupervised acoustic models for speech recognition and keyword spotting (ISCA, Baixas, France, 2014), pp. 16–20
83. S Young, G Evermann, M Gales, T Hain, D Kershaw, X Liu, G Moore, J Odell, D Ollason, D Povey, V Valtchev, P Woodland, *The HTK Book*. (Engineering Department, Cambridge University, 2006)
84. L Mangu, H Soltau, H-K Kuo, B Kingsbury, G Saon, in *Proc. of ICASSP*. Exploiting diversity for spoken term detection (Vancouver, BC, 2013), pp. 8282–8286
85. V-B Le, L Lamel, A Messaoudi, W Hartmann, J-L Gauvain, C Woehrling, J Despres, A Roy, in *Proc. of Interspeech*. Developing STT and KWS systems using limited language resources (ISCA, Baixas, France, 2014), pp. 2484–2488
86. H-Y Lee, Y Zhang, E Chuangsuwanich, J Glass, in *Proc. of Interspeech*. Graph-based re-ranking using acoustic feature similarity between search results for spoken term detection on low-resource languages (ISCA, Baixas, France, 2014), pp. 2479–2483
87. M Ma, J Richards, V Soto, J Hirschberg, A Rosenberg, in *Proc. of Interspeech*. Strategies for rescoring keyword search results using word-burst and acoustic features (ISCA, Baixas, France, 2014), pp. 2769–2773
88. J Chiu, Y Wang, J Trmal, D Povey, G Chen, A Rudnicky, in *Proc. of Interspeech*. Combination of FST and CN search in spoken term detection (ISCA, Baixas, France, 2014), pp. 2784–2788
89. VT Pham, H Xu, NF Chen, S Sivadas, BP Lim, ES Chng, H Li, in *Proc. of ICASSP*. Discriminative score normalization for keyword search decision (Florence, 2014), pp. 7078–7082
90. S Wegmann, A Faria, A Janin, K Riedhammer, N Morgan, in *Proc. of ASRU*. The TAO of ATWV: Probing the mysteries of keyword search performance (Olomouc, 2013), pp. 192–197
91. D Karakos, R Schwartz, S Tsakalidis, L Zhang, S Ranjan, T Ng, R Hsiao, G Saikumar, I Bulyko, L Nguyen, J Makhoul, F Grezl, M Hannemann, M Karafiat, I Szoke, K Vesely, L Lamel, V-B Le, in *Proc. of ASRU*. Score normalization and system combination for keyword spotting (Olomouc, 2013), pp. 210–215
92. Z Chen, T Zhang, J Wu, in *Proc. of SLT*. Subword scheme for keyword search (South Lake Tahoe, NV, 2014), pp. 483–488
93. M Bisani, H Ney, Joint-sequence models for grapheme-to-phoneme conversion. Speech Commun. **50**(5), 434–451 (2008)
94. J Trmal, G Chen, D Povey, S Khudanpur, P Ghahremani, X Zhang, V Manohar, C Liu, A Jansen, D Klakow, D Yarowsky, F Metze, in *Proc. of SLT*. A keyword search system using open source software (IEEE, USA, 2014), pp. 530–535
95. A Jansen, P Niyogi, Point process models for spotting keywords in continuous speech. IEEE Trans. Audio, Speech Lang. Process. **17**(8), 1457–1470 (2009)
96. K Kintzley, A Jansen, H Hermansky, in *Proc. of ICASSP*. Featherweight phonetic keyword search for conversational speech (Florence, 2014), pp. 7859–7863
97. B Zhang, R Schwartz, S Tsakalidis, L Nguyen, S Matsoukas, in *Proc. of Interspeech*. White listing and score normalization for keyword spotting of noisy speech (ISCA, Baixas, France, 2012), pp. 1832–1835
98. A Mandal, J van Hout, Y-C Tam, V Mitra, Y Lei, J Zheng, D Vergyri, L Ferrer, M Graciarena, A Kathol, H Franco, in *Proc. of Interspeech*. Strategies for high accuracy keyword detection in noisy channels (Lyon, France, 2013), pp. 15–19
99. T Ng, R Hsiao, L Zhang, D Karakos, SH Mallidi, M Karafiat, K Vesely, I Szoke, B Zhang, L Nguyen, R Schwartz, in *Proc. of Interspeech*. Progress in the BBN keyword search system for the DARPA RATS program (ISCA, Baixas, France, 2014), pp. 959–963
100. J van Hout, V Mitra, Y Lei, D Vergyri, M Graciarena, A Mandal, H Franco, in *Proc. of Interspeech*. Recent improvements in SRI's keyword detection system for noisy audio (ISCA, Baixas, France, 2014), pp. 1727–1731
101. L Mangu, H Soltau, H-K Kuo, G Saon, in *Proc. of ASRU*. The IBM keyword se arch system for the DARPA RATS program (Olomouc, 2013), pp. 204–209
102. MS Seigel, PC Woodland, MJF Gales, in *Proc. of ICASSP*. A confidence-based approach for improving keyword hypothesis scores (IEEE, USA, 2013), pp. 8565–8569

103. V Mitra, J van Hout, H Franco, D Vergyri, Y Lei, M Graciarena, Y-C Tam, J Zheng, in *Proc. of ICASSP*. Feature fusion for high-accuracy keyword spotting (IEEE, USA, 2014), pp. 7143–7147
104. A Martin, G Doddington, T Kamm, M Ordowski, M Przybocki, in *Proc. of Eurospeech*. The DET curve in assessment of detection task performance (ISCA, Baixas, France, 1997), pp. 1895–1898
105. Consorcio MAVIR. http://www.mavir.net
106. SoX - Sound eXchange. http://sox.sourceforge.net/
107. NIST, *NIST Speech Tools and APIs: 2006*. (National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 1996). National Institute of Standards and Technology (NIST). http://www.nist.gov/speech/tools/index.htm
108. Vídeos corpus MAVIR. http://cartago.lllf.uam.es/mavir/index.pl?m=videos
109. Spoken Term Detection Evaluation (2006). http://www.itl.nist.gov/iad/mig/tests/std/2006/
110. NF Chen, JG Fiscus, in *Speech and Language Technical Committee Newsletter*. Overview of the NIST Open Keyword Search 2013 Evaluation Workshop (IEEE, USA, 2013)
111. M Cai, Z Lv, B Song, Y Shi, W Wu, C Lu, W-Q Zhang, J Liu, in *Proc. of ICASSP*. The THUEE system for the OpenKWS14 keyword search evaluation (IEEE, USA, 2015), pp. 4734–4738
112. H Su, VT Pham, Y He, J Hieronymus, in *Proc. of ICASSP*. Improvements on transducing syllable lattice to word lattice for keyword search (IEEE, USA, 2015), pp. 4729–4733
113. F Metze, A Gandhe, Y Miao, Z Sheikh, Y Wang, D Xu, H Zhang, J Kim, I Lane, WK Lee, S Stuker, M Muller, in *Proc. of ICASSP*. Semi-supervised training in low-resource ASR and KWS (IEEE, USA, 2015), pp. 4699–4703
114. C Ni, C-C Leung, L Wang, NF Chen, B Ma, in *Proc. of ICASSP*. Unsupervised data selection and word-morph mixed language model for Tamil low-resource keyword search (IEEE, USA, 2015), pp. 4714–4718
115. NF Chen, C Ni, I-F Chen, S Sivadas, VT Pham, H Xu, X Xiao, TS Lau, SJ Leow, BP Lim, C-C Leung, L Wang, C-H Lee, A Goh, ES Chng, B Ma, H Li, in *Proc. of ICASSP*. Low-resource keyword search strategies for Tamil (IEEE, USA, 2015), pp. 5366–5370
116. L Mangu, G Saon, M Picheny, B Kingsbury, in *Proc. of ICASSP*. Order-free spoken term detection (IEEE, USA, 2015), pp. 5331–5335
117. I-F Chen, C Ni, BP Lim, NF Chen, C-H Lee, in *Proc. of ICASSP*. A keyword-aware grammar framework for LVCSR-based spoken keyword search (IEEE, USA, 2015), pp. 5196–5200
118. J Garofolo, G Auzanne, E Voorhees, in *Proc. of the Text Retrieval Conference (TREC-8)*. The TREC spoken document retrieval track: A success story (NIST, Gaithersburg, MD, USA, 2000)
119. B Logan, P Moreno, JMV Thong, E Whittaker, in *Proc. of ICSLP*. An experimental study of an audio indexing system for the web (ISCA, Baixas, France, 2000), pp. 676–679
120. A Moreno-Sandoval, DT Toledano, R de la Torre, M Garrote, JM Guirao, in *Proc. of LREC*. Developing a phonemic and syllabic frequency inventory for spontaneous spoken castilian spanish and their comparison to text-based inventories (ELRA, Paris, France, 2008), pp. 1097–1100
121. Trusted Connections. http://www.tc-star.org
122. L Docío-Fernández, A Cardenal-López, C García-Mateo, in *Proc. of TC-STAR Workshop on Speech-to-Speech Translation*. TC-STAR 2006 automatic speech recognition evaluation: The uvigo system (ELRA, Paris, France, 2006)
123. Consulta del corpus MAVIR. http://cartago.lllf.uam.es/mavir/index.pl?m=descargas
124. D Can, M Saraclar, Lattice indexing for spoken term detection. IEEE Trans. Audio Speech Lang. Process. **19**(8), 2338–2347 (2011)
125. A Cardenal-Lopez, C Garcia-Mateo, A Espiña, P Lopez-Otero, L Docio-Fernandez, in *Proc. of IberSpeech*. GTMTranscriber: a subtitling program for broadcast news in Galician language (Springer, United Kingdom, 2012), pp. 377–382
126. A Cardenal-López, FJ Diéguez-Tirado, C García-Mateo, in *Proc. of ICASSP*. Fast LM Look-Ahead for Large Vocabulary Continuous Speech Recognition Using Perfect Hashing (IEEE, USA, 2002), pp. 705–708
127. A Stolcke, in *Proc. of Interspeech*. SRILM - an extensible language modeling toolkit (ISCA, Baixas, France, 2002), pp. 901–904
128. A Abad, LJ Rodríguez-Fuentes, M Peñagarikano, A Varona, G Bordel, in *Proc. of Interspeech*. On the calibration and fusion of heterogeneous spoken term detection systems (ISCA, Baixas, France, 2013), pp. 20–24
129. BOSARIS Toolkit. https://sites.google.com/site/bosaristoolkit/
130. VOICEBOX: Speech Processing Toolbox for MATLAB. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
131. D Mostefa, O Hamon, N Moreau, K Choukri, Evaluation report for the technology and corpora for speech to speech translation. TC-STAR Project. Deliverable N. 30 (2007). http://www.tcstar.org/documents/D30.pdf
132. A Moreno, D Poch, A Bonafonte, E Lleida, J Llisterri, J Mariño, C Nadeu, in *Proc. of Eurospeech*. ALBAYZIN speech database: Design of the phonetic corpus (ISCA, Baixas, France, 1993), pp. 653–656
133. A Moreno, B Lindberg, C Draxler, in *Proc. of LREC*. Speechdat-car: A large speech database for automotive environments (ELRA, Paris, France, 2000)
134. R Justo, O Saz, V Guijarrubia, A Miguel, MI Torres, E Lleida, in *Proc. of International Conference on Ambient Media and Systems*. Improving dialogue systems in a home automation environment (ACM, New York, USA, 2008), pp. 1–6
135. HVD Heuvel, K Choukri, C Gollan, in *Proc. of LREC*. TC-STAR: New language resources for ASR and SLT purposes (ELRA, Paris, France, 2006), pp. 2570–2573
136. CJ Chen, RA Gopinath, MD Monkowski, MA Picheny, K Shen, in *Proc. of Eurospeech*. New methods in continuous Mandarin speech recognition (ISCA, Baixas, France, 1997), pp. 22–25
137. AM Azmi, RS Almajed, A survey of automatic Arabic diacritization techniques. Nat. Lang. Eng. **21**(3), 477–495 (2013)