Tejedor *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2016) 2016:1
DOI 10.1186/s13636-016-0080-2

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

RESEARCH Open Access

Comparison of ALBAYZIN query-by-example spoken term detection 2012 and 2014 evaluations

Javier Tejedor[1]*, Doroteo T. Toledano[2], Paula Lopez-Otero[3], Laura Docio-Fernandez[3] and Carmen Garcia-Mateo[3]

Abstract

Query-by-example spoken term detection (QbE STD) aims at retrieving data from a speech repository given an acoustic query containing the term of interest as input. Nowadays, it is receiving much interest due to the large volume of multimedia information. This paper presents the systems submitted to the ALBAYZIN QbE STD 2014 evaluation held as a part of the ALBAYZIN 2014 Evaluation campaign within the context of the IberSPEECH 2014 conference. This is the second QbE STD evaluation in Spanish, which allows us to evaluate the progress in this technology for this language. The evaluation consists in retrieving the speech files that contain the input queries, indicating the start and end times where the input queries were found, along with a score value that reflects the confidence given to the detection of the query. Evaluation is conducted on a Spanish spontaneous speech database containing a set of talks from workshops, which amount to about 7 h of speech. We present the database, the evaluation metric, the systems submitted to the evaluation, the results, and compare this second evaluation with the first ALBAYZIN QbE STD evaluation held in 2012. Four different research groups took part in the evaluations held in 2012 and 2014. In 2014, new multi-word and foreign queries were added to the single-word and in-language queries used in 2012. Systems submitted to the second evaluation are hybrid systems which integrate letter transcription- and template matching-based systems. Despite the significant improvement obtained by the systems submitted to this second evaluation compared to those of the first evaluation, results still show the difficulty of this task and indicate that there is still room for improvement.

Keywords: Query-by-example spoken term detection, International evaluation, Search on spontaneous speech

Introduction

The ever-increasing volume of heterogeneous speech data stored in audio and audiovisual repositories promotes the development of efficient methods for retrieving such information. Significant research has been conducted on spoken document retrieval (SDR), keyword spotting (KWS), spoken term detection (STD), query-by-example (QbE), or spoken query approaches to address this issue. STD aims at finding individual words or sequences of words within audio archives. It usually relies on a text-based input, commonly the word/phone transcription of the search term. For this reason, STD is also called text-based STD. STD systems are typically composed of three different stages: (1) the audio is decoded in terms of word/subword lattices using an automatic speech recognition (ASR) subsystem, (2) a term detection subsystem searches the terms within those word/subword lattices and hypothesizes detections, and (3) confidence measures are applied to output reliable detections.

QbE can be defined as 'a method of searching for an example of an object or a part of it in other objects.' This has been widely used in audio applications such as sound classification [1–3], music information retrieval [4, 5], and spoken document retrieval [6]. In QbE STD, we consider the scenario in which the user has found some interesting data within a speech data repository and his/her purpose is to find similar data within that repository. The interesting data consist of one or several speech segments containing the term of interest (henceforth,

*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

query) and the system outputs other putative hits from the repository (henceforth, utterances). Alternatively, the term of interest can be uttered by the user. Using speech queries offers a big advantage for devices with limited text-based capabilities, which can be effectively used under the QbE STD paradigm. Other advantage is that QbE STD can be employed for building language-independent STD systems [7–10], since prior knowledge of the language involved in the speech data is not necessary.

QbE STD has been mainly addressed in the literature from two perspectives:

- Methods based on word/subword transcription of the query [11–17], so that the text-based STD technology can be applied.
- Methods based on template matching of features extracted from the query and speech repository [9–11, 15, 18–27]. These usually borrow the idea from dynamic time warping (DTW)-based speech recognition and were found to outperform subword transcription-based techniques in QbE STD [28].

Recently, hybrid systems combining both methods have also been proposed [29–33].

Unsupervised spoken term detection techniques, which aim at automatically discovering acoustic patterns (e.g., for training acoustic models) for languages for which manual transcriptions and linguistic knowledge are scarce, have been also investigated [34, 35]. These techniques can also be employed for building language-independent QbE STD systems, since prior knowledge of the language is not necessary.

Recently, several evaluations including SDR, STD, and QbE STD have been held [36–40]. We organized the second international ALBAYZIN QbE STD evaluation in the context of the ALBAYZIN 2014 Evaluation campaign. These campaigns are internationally open sets of evaluations supported by the Spanish Network of Speech Technologies (RTTH)[1] and the ISCA Special Interest Group on Iberian Languages (SIG-IL)[2] held every 2 years from 2006. The evaluation campaigns provide an objective mechanism to compare different systems and promote research in different speech technologies such as audio segmentation [41], speaker diarization [42], language recognition [43], query-by-example spoken term detection [44], and speech synthesis [45].

Spanish is a major language in the world and significant research has been conducted on it for ASR, KWS, and STD tasks [46–52]. In 2012, the first QbE STD evaluation dealing with Spanish was organized in the context of the ALBAYZIN 2012 Evaluation campaign. The success of this first evaluation [44] encouraged us to organize a new QbE STD evaluation for the ALBAYZIN 2014 Evaluation campaign aiming at evaluating the progress in this technology for Spanish. The second ALBAYZIN QbE STD evaluation incorporates new and more difficult queries (i.e., multi-word and foreign queries). In addition, all the queries from the first evaluation are kept so that a comparison between the systems submitted to both evaluations is possible. This paper presents the systems submitted to the ALBAYZIN QbE STD 2014 evaluation and makes a comparison with the systems submitted to the ALBAYZIN QbE STD 2012 evaluation.

The rest of the paper is organized as follows: The next section presents a description of the QbE STD evaluations held in 2012 and 2014. Section 3 presents the different systems submitted to the evaluations. Results along with discussion are presented next, and the work is concluded in the last section.

ALBAYZIN QbE STD evaluations

The ALBAYZIN QbE STD 2012 and 2014 evaluations involve searching for audio content within audio content using an audio content query. These evaluations are suitable for research groups working on speech indexing and retrieval and on speech recognition. The input to the system is an acoustic example per query, and hence prior knowledge of the correct word/subword transcription corresponding to each query is not available.

The evaluations consist in searching for a test query list within test speech data. Participants were provided with a training/development (train/dev) query list and speech data that can be used for system training and tuning, though any additional data can also be employed, as long as it is properly described in the system description.

Participants could submit a primary system and several contrastive systems. No manual intervention is allowed to generate the final output file, and hence all 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 on test data have been sent to the participants. The standard XML-based format corresponding to the NIST STD 2006 evaluation [53] has been used for building the system output file. For both evaluations, about 3 months were given to the participants for system design. Train/dev data (i.e., train/dev speech data, train/dev query list, train/dev ground-truth labels, orthographic transcription and timestamps for phrase boundaries in the train/dev speech data, and evaluation tools) were released at the end of June 2012/2014, test data (i.e., test speech data and test query list) were released at the beginning of September 2012/2014, and the final system submission was due at the end of September 2012/2014. Final results were presented and discussed at IberSPEECH 2012 and IberSPEECH 2014 conferences at the end of November 2012[3]/2014[4].

Evaluation metric

In QbE STD, a hypothesized occurrence is called a *detection*; if the detection corresponds to an actual occurrence, it is called a *hit*, otherwise it is a *false alarm* (FA). An actual occurrence that is not detected is called a *miss*. The Actual Term Weighted Value (ATWV) proposed by NIST [53] for STD has been used as the main metric for the evaluations. This metric integrates the hit rate and false alarm rate of each query into a single metric and then averages over all the queries:

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

where Δ denotes the set of queries and $|\Delta|$ is the number of queries in this set. N^K_{hit} and N^K_{FA} represent the numbers of hits and false alarms of query K, respectively, and N^K_{true} 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 [54]. 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 system (usually tuned on development data). An additional metric, called Maximum Term Weighted Value (MTWV) [53] has also been considered. It is the maximum TWV achieved by the system for all possible thresholds and hence does not depend on the tuned threshold. This MTWV represents the performance of the system if the threshold was perfectly set. Results based on this metric are presented to evaluate the goodness of threshold selection.

In addition to ATWV and MTWV, NIST also proposed a detection error tradeoff (DET) curve [55] to evaluate the performance of a QbE STD system working at various miss/FA ratios. DET curves are also presented in this paper for system comparison.

Database

The database used for ALBAYZIN QbE STD 2012 and 2014 evaluations consists of a set of talks extracted from the MAVIR workshops[5] 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 evaluations. 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 these evaluations into train/dev and test sets. There are 20 male and 3 female speakers in the MAVIR Spanish database.

The speech data were originally recorded in several audio formats (pulse-code modulation (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[6]. Recordings were made with the same equipment, a Digital TASCAM DAT model DA-P1, except for one recording. Different microphones were used for the different recordings. They mainly consisted of tabletop or floor standing microphones, but in one case a lavalier microphone was used. The distance from the mouth of the speaker to the microphone varies and was not particularly controlled, but in most cases the distance was smaller than 50 cm. All the 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 ALBAYZIN QbE STD evaluations and definitely for further work. The speech data were manually annotated in an orthographic form, but timestamps were only set for phrase boundaries. To prepare the data used for the evaluations, organizers manually added the timestamps for all the occurrences of the train/dev and test search queries. Table 1 includes some database features such as the number of word occurrences, duration, and signal-to-noise ratio (SNR) [56] of each speech file in the MAVIR Spanish database.

Given that the database used in these evaluations consists of spontaneous speech, there is an inherent difficulty for query detection. In addition, QbE STD and, in general, ASR system performance significantly degrades when training data belong to a different domain or pose different acoustic conditions to those of the test data. To alleviate this problem in the ALBAYZIN QbE STD evaluations, organizers provided limited train/dev data from the same domain and with *similar* acoustic conditions (microphone speech from workshops) as the test data. However, it must be noted that different microphones and even different rooms were used for each recorded file in the MAVIR database, and hence, the acoustic conditions can vary significantly from one file to another.

Train/dev 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. For the evaluation held in 2012, the train/dev query list consisted of 60 queries. All of them were single words with lengths between 7 and 16 graphemes. There are 1027 occurrences of those queries in the train/dev speech data. For the evaluation held in 2014, the train/dev

Table 1 MAVIR database characteristics

File ID	Dataset	# word occ.	Duration (min)	# speakers	SNR (dB)
Mavir-02	Train/dev	13,432	74.51	7 male	2.1
Mavir-03	Train/dev	6681	38.18	1 male, 1 fem.	15.8
Mavir-06	Train/dev	4332	29.15	2 male, 1 fem.	12.0
Mavir-07	Train/dev	3831	21.78	2 male	10.6
Mavir-08	Train/dev	3356	18.90	1 male	7.5
Mavir-09	Train/dev	11,179	70.05	1 male	12.3
Mavir-12	Train/dev	11,168	67.66	1 male	11.1
Mavir-04	Test	9310	57.36	3 male, 1 fem.	10.2
Mavir-11	Test	3130	20.33	1 male	9.2
Mavir-13	Test	7837	43.61	1 male	11.1
All	Train/dev	53,979	320.23	15 male and 2 fem.	–
All	Test	20,277	121.3	5 male and 1 fem.	–

occ. occurrences, *min* minutes, *fem.* female, *SNR* signal-to-noise ratio, *dB* decibels

query list consists of 94 queries. Each query is composed of one or more words containing between 5 and 18 graphemes. There are 1415 occurrences of those queries in the train/dev speech data. Tables 2 and 3 include information related to the train/dev queries used in both 2012 and 2014 evaluations and the train/dev queries used only in the 2014 evaluation, respectively.

Test data amount to about 2 h of speech extracted from the other 3 speech files not included in train/dev data and contain 5 male and 1 female speakers. For the evaluation held in 2012, the test query list consisted of 60 queries. All of them were single words containing between 7 and 16 graphemes. There are 892 occurrences of those queries in the test speech data. For the evaluation held in 2014, the

Table 2 Training/development queries used in both 2012 and 2014 evaluations. Each cell indicates query text, time length per query (in milliseconds), and number of occurrences per query

Query (time)–(# occ.)	Query (time)–(# occ.)	Query (time)–(# occ.)
Académico (500)–(10)	Gallego (300)–(7)	Cuestión (260)–(8)
Acceder (350)–(7)	General (350)–(43)	Cultural (790)–(10)
Administración (550)–(27)	Indexación (640)–(10)	Desarrollo (750)–(15)
Arquitectura (610)–(8)	Industria (390)–(6)	Después (280)–(38)
Barcelona (670)–(8)	Información (570)–(153)	Directamente (450)–(16)
Cálculo (440)–(6)	Instituto (370)–(22)	Establecer (550)–(8)
Calidad (550)–(33)	Investigación (740)–(52)	Estructura (540)–(13)
Capacidad (670)–(12)	Latinoamérica (690)–(8)	Euskera (530)–(10)
Capital (500)–(11)	Máquina (510)–(8)	Formato (430)–(7)
Castellano (670)–(21)	Ministerio (310)–(9)	Francia (560)–(6)
Catalogación (750)–(6)	Momento (370)–(50)	Sentido (380)–(24)
Cataluña (440)–(11)	Nacional (770)–(7)	Situación (690)–(24)
Cervantes (420)–(25)	Negocio (490)–(18)	Soporte (330)–(6)
Clasificación (620)–(13)	Patrimonio (670)–(7)	Telefónica (540)–(21)
Comentario (540)–(14)	Pequeño (320)–(8)	Todavía (330)–(16)
Compañía (360)–(6)	Validación (520)–(7)	Publicidad (650)–(13)
Picasso (270)–(21)	Conjunto (340)–(16)	Visibilidad (730)–(8)
Trabajo (320)–(36)	Proceso (420)–(13)	Contabilidad (1090)–(7)
Computadora (740)–(12)	Virtual (570)–(12)	Referencia (530)–(9)
Potencial (470)–(13)	Conocimiento (560)–(6)	Volumen (300)–(6)

Table 3 New training/development queries used in the 2014 evaluation and not in the 2012 evaluation, time length per query (in milliseconds), and number of occurrences per query

Query (time)–(# occ.)	Query (time)–(# occ.)
Presentación (760)–(17)	Portugal (340)–(4)
Vosotros (360)–(6)	Parlamento (360)–(3)
Etcétera (490)–(28)	Microsoft (580)–(4)
Empresas (820)–(71)	Mavir (420)–(2)
Porcentaje (490)–(6)	Málaga (450)–(2)
Experimentos (400)–(10)	Isabel (310)–(4)
Noventa (630)–(39)	Garner (320)–(3)
Atención (280)–(8)	Galicia (520)–(4)
Mercado (510)–(111)	Erasmus (430)–(2)
Resolver (500)–(8)	Dilbert (640)–(2)
Probablemente (490)–(6)	Complutense (460)–(4)
Dominios (370)–(17)	Cristian (510)–(2)
Wikipedia (670)–(3)	Berrilan (430)–(2)
Webmaster (460)–(2)	Aguilera (480)–(3)
Valladolid (530)–(2)	Premios nobel (650)–(2)
Sevilla (450)–(2)	Universidad de Chile (840)–(3)
Profit (330)–(3)	Nick cohn (410)–(3)

test query list consists of 99 queries. Each query is composed of one or more words containing between 6 and 16 graphemes. There are 1162 occurrences of those queries in the test speech data. Tables 4 and 5 include information related to the test queries used in both 2012 and 2014 evaluations and the test queries used only in the 2014 evaluation, respectively.

Each train/dev query has one or more occurrences in the train/dev speech data and each test query has one or more occurrences in the test speech data. All these queries were extracted from the MAVIR database.

Comparison with other QbE STD evaluations

The most similar evaluations to ALBAYZIN QbE STD evaluations are the MediaEval 2011, 2012, and 2013 Spoken Web Search [38, 57, 58]. The task to be performed in MediaEval and ALBAYZIN evaluations is the same, but these differ in several aspects. This makes it difficult to compare the results obtained in ALBAYZIN evaluations to previous MediaEval evaluations.

The most important difference is the nature of the audio content used for the evaluations. In MediaEval evaluations, the speech is typically telephone speech, either conversational or read and elicited speech, or speech recorded with in-room microphones. In ALBAYZIN evaluations, the audio contains microphone recordings of real talks in real workshops, in large conference rooms with the public. Microphones, conference rooms, and even recording conditions change from one recording to another. Microphones are not close talking microphones but table top and floor standing microphones mainly.

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

Besides MediaEval evaluations, a new QbE STD evaluation has been organized within NTCIR-11 conference [59]. Data used in this evaluation contained spontaneous speech in Japanese provided by the National Institute for Japanese language as well as spontaneous speech recorded during seven editions of the Spoken Document Processing Workshop. As additional information, this evaluation provides participants with the results of a voice activity detection system on the input speech data, the manual transcription of the speech data, and the output of a Large Vocabulary Continuous Speech Recognition (LVCSR) system. Although ALBAYZIN evaluations could be similar in terms of speech nature to this NTCIR QbE STD evaluation (speech recorded in real workshops), ALBAYZIN evaluations do not provide any kind of information apart from the speech content, the list of queries, and the train/dev ground-truth files to participants. In addition, ALBAYZIN evaluations make use of other language and define disjoint train/dev and test query lists to measure the generalization capability of the systems.

Table 6 summarizes the main characteristics of the MediaEval QbE STD evaluations, the NTCIR-11 QbE STD evaluation, and the ALBAYZIN QbE STD evaluations held in 2012 and 2014.

Systems

Four research groups, listed in Table 7, took part in the ALBAYZIN QbE STD evaluations held in 2012 and 2014. Four systems were submitted to the ALBAYZIN QbE STD 2014 evaluation. Two of them (those based on deep neural networks (DNN)) were post-evaluation submissions. In addition, two text-based STD systems (the DNN-based system was a post-evaluation submission) were also submitted to compare the QbE STD systems with other technology that also aims at searching for terms within speech data. For the ALBAYZIN QbE STD 2012 evaluation, four different systems were submitted. Table 8 summarizes the main characteristics of these QbE STD systems, which are further described next along with the text-based STD system.

ALBAYZIN QbE STD 2014 evaluation: Fusion (SGMM)+Posteriorgram system

This system consists of the fusion of three different subsystems, as shown in Fig. 1: a large vocabulary continuous speech recognition system, a dynamic time warping

Table 4 Test queries used in both 2012 and 2014 evaluations. Each cell indicates query text, time length per query (in milliseconds), and number of occurrences per query

Query (tme)–(# occ.)	Query (time)–(# occ.)	Query (time)–(# occ.)
Acuerdo (290)–(7)	Lenguaje (390)–(6)	Referencia (470)–(13)
Análisis (370)–(18)	Mecanismo (470)–(7)	Fuenlabrada (570)–(15)
Aproximación (850)–(7)	Metodología (810)–(10)	General (420)–(11)
Buscador (580)–(7)	Motores (340)–(6)	Gracias (400)–(13)
Cangrejo (490)–(7)	Necesario (650)–(6)	Idiomas (290)–(27)
Castellano (570)–(9)	Normalmente (320)–(6)	Implicación (600)–(31)
Conjunto (490)–(7)	Obtener (380)–(9)	Importante (680)–(19)
Conocimiento (490)–(6)	Orientación (600)–(6)	Incluso (410)–(12)
Desarrollo (460)–(6)	Parecido (400)–(6)	Información (560)–(92)
Detalle (280)–(7)	Personas (540)–(6)	Intentar (420)–(13)
Difícil (410)–(12)	Perspectiva (490)–(7)	Interfaz (480)–(10)
Distintos (450)–(21)	Porcentaje (660)–(8)	Resolver (420)–(6)
Documentos (750)–(7)	Precisamente (680)–(6)	Segunda (520)–(8)
Efectivamente (290)–(10)	Presentación (580)–(15)	Seguridad (350)–(6)
Ejemplo (550)–(54)	Primera (290)–(19)	Siguiente (370)–(11)
Empezar (340)–(7)	También (240)–(93)	Reconocimiento (660)–(6)
Principio (480)–(9)	Entidades (670)–(28)	Trabajar (380)–(39)
Simplemente (650)–(8)	Realidad (270)–(10)	Evaluación (480)–(15)
Encontrar (350)–(19)	Textual (590)–(15)	Recurso (520)–(7)
Propuesta (440)–(19)	Estudiar (500)–(7)	Utilizar (500)–(15)

search-based system with a fingerprint representation of the queries and the utterances, and a DTW search-based system with phoneme posterior probabilities for query and utterance representation. The three subsystems are described in the following sections.

Kaldi LVCSR-based QbE STD system

The architecture of the Kaldi LVCSR-based QbE STD system is shown in Fig. 2. First, an LVCSR system was built using the Kaldi open-source toolkit [60]. Thirteen-dimensional perceptual linear prediction (PLP)

Table 5 New test queries used in the 2014 evaluation and not in the 2012 evaluation, time length per query (in milliseconds), and number of occurrences per query

Query (time)–(# occ.)	Query (time)–(# occ.)	Query (time)–(# occ.)
Académico (690)–(6)	Investigación (520)–(15)	Formularios (520)–(19)
Anselmo (440)–(2)	Madrid (410)–(6)	Transparencia (500)–(6)
Antonio moreno (480)–(2)	Manuel (240)–(6)	Francisco garcía (930)–(2)
Autónoma (690)–(8)	Mencionado (500)–(6)	Unesco (390)–(5)
Bastante (590)–(22)	Objetivos (650)–(9)	Fundamentalmente (900)–(7)
Cindoc (690)–(2)	Obviamente (550)–(10)	Universidades (790)–(13)
Coca cola (470)–(2)	Paloma (310)–(3)	Harvard (540)–(2)
Completamente (450)–(13)	Programa (420)–(11)	Validación (500)–(18)
Consorcio mavir (820)–(2)	Scholar (390)–(2)	Huelva (250)–(2)
Daedalus (660)–(6)	Setenta (430)–(6)	Vicente fox (730)–(2)
Embargo (340)–(7)	Solamente (620)–(10)	Inicial (590)–(8)
Evidentemente (660)–(7)	Solución (560)–(7)	Zagreb (460)–(2)
Felisa (310)–(2)	Soporte (600)–(89)	Internet (410)–(4)

Table 6 QbE STD evaluation characteristics and languages: Albanian ('ALB'), Basque ('BAS'), Czech ('CZE'), non-native English ('NN-ENG'), Isixhosa ('ISIX'), Isizulu ('ISIZ'), Romanian ('ROM'), Sepedi ('SEP'), Setswana ('SET'), and Slovak ('SLO')

Evaluation	Language/s	Type of speech	# queries dev./test	Metric
MediaEval 2011	English, Hindi, Gujarati, and Telugu	Tel.	64/36	ATWV
MediaEval 2012	2011 + isiNdebele, Siswati, Tshivenda, and Xitsonga	Tel.	164/136	ATWV
MediaEval 2013	ALB, BAS, CZE, NN-ENG, ISIX, ISIZ, ROM, SEP, and SET	Tel. and mic.	>600/>600	ATWV
MediaEval 2014	ALB, BAS, CZE, NN-ENG, ROM, and SLO	Tel. and mic.	560/555	C_{nxe}
NTCIR-11 2014	Japanese	mic. workshop	63/203	F-measure
ALBAYZIN 2012	Spanish	mic. workshop	60/60	ATWV
ALBAYZIN 2014	Spanish	mic. workshop	94/99	ATWV

Tel. telephone, *mic.* microphone, *dev.* development, C_{nxe} normalized cross entropy cost

coefficients augmented with delta and double delta coefficients were used to build 39-dimensional feature vectors used as acoustic features. Next, 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 Hidden Markov Models (HMMs) and ends with speaker adaptive training (SAT) of state-clustered triphone HMMs with Gaussian mixture model (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 a Subspace GMM (SGMM) employed in the decoding stage to generate word lattices and word sequences.

Table 7 Participants in the ALBAYZIN QbE STD 2012 and 2014 evaluations along with their submitted systems

Team ID	Research institution	Year	System/s
TID	Telefonica Research, Barcelona, Spain	2012	DTW-Zero
GTTS	University of the Basque Country, Bilbao, Spain	2012	P1B-STD
ELiRF	Politechnical University of Valencia, Valencia, Spain	2012	P1L-STD DTW-Spanish
GTM	AtlantTIC Research Center, University of Vigo, Vigo, Spain	2014	Fusion+Post. Fusion

Post. posteriorgram

Table 8 Main characteristics of the ALBAYZIN QbE STD 2012 and 2014 evaluation systems

System ID	System type	Language-dependent
Fusion (SGMM)+Posteriorgram	Fusion: 1 SGMM-based LVCSR and 2 template matching	YES
Fusion (SGMM)	Fusion: 1 SGMM-based LVCSR and 1 template matching	YES
Fusion (DNN)+Posteriorgram	Fusion: 1 DNN-based LVCSR and 2 template matching	YES
Fusion (DNN)	Fusion: 1 DNN-based LVCSR and 1 template matching	YES
DTW-Zero	Template matching	NO
P1B-STD	Phone transcription	NO
P1L-STD	Phone transcription	YES
DTW-Spanish	Template matching	YES

LVCSR Large Vocabulary Continuous Speech Recognition

The acoustic models were trained using the Spanish data from 2006 TC-STAR ASR evaluation campaign[7]. Specifically, the training data from the European Parliamentary plenary sessions and the Spanish Parliament sessions, which were manually transcribed, were used for acoustic model training [61]. 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. After this, 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 European and Spanish Parliaments of the TC-STAR database, subtitles, books, newspapers, online courses, and the transcriptions of the MAVIR sessions included in the train/dev data provided by the organizers[8]. 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 train/dev data files except one were then used to compute the query 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 60,000 words and a Kneser-Ney discount strategy was

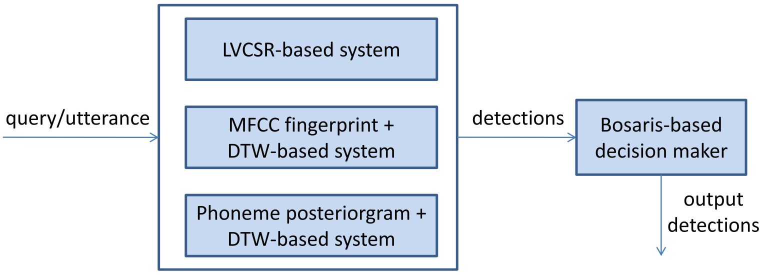
LVCSR-based system
MFCC fingerprint + DTW-based system
Phoneme posteriorgram + DTW-based system
query/utterance
detections
Bosaris-based decision maker
output detections

Fig. 1 Fusion (SGMM)+Posteriorgram system architecture

used for decoding. LMs have been built using the SRILM toolkit [62].

The Kaldi-based LVCSR system generates word lattices [63] for the utterances and a word sequence for each query using the SGMM acoustic models and the word-based LM. The word sequence of the query is then used as the word transcription of the query term so that a text-based STD system can be effectively used to hypothesize query detections. To do so, the system integrates the Kaldi term detector [60, 64, 65], which searches for the input queries within those word lattices. The lattice indexing technique, described in [66], first converts the word lattices of all the utterances from individual weighted finite state transducers (WFST) 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 3-dimensional cost. This factor transducer represents an inverted index of all the word sequences contained in the lattices. Thus, given a search query, a simple finite state machine that accepts the input query (from its word sequence) is created and composed with the factor transducer in order to obtain all the occurrences of the query in the utterances. The posterior probabilities of the lattice corresponding to all the words of the input query are accumulated, assigning a confidence score to each detection.

MFCC fingerprint-based QbE STD system

This system employs a Mel-frequency cepstral coefficient (MFCC) fingerprint representation of the queries and utterances [67], and search is performed using a dynamic time warping approach.

An audio fingerprint is a compact representation of an audio signal that extracts the most meaningful information of an audio excerpt. This representation, which is often restricted to binary values, has been used in different tasks [68–70] for several reasons: (1) the storage requirements of fingerprints are relatively small since binary values are employed, (2) the comparison of different fingerprints is efficient since perceptual irrelevancies have been removed, and (3) searching on fingerprint databases is efficient because the searching space is small [71] and almost all the operations are performed in a binary domain.

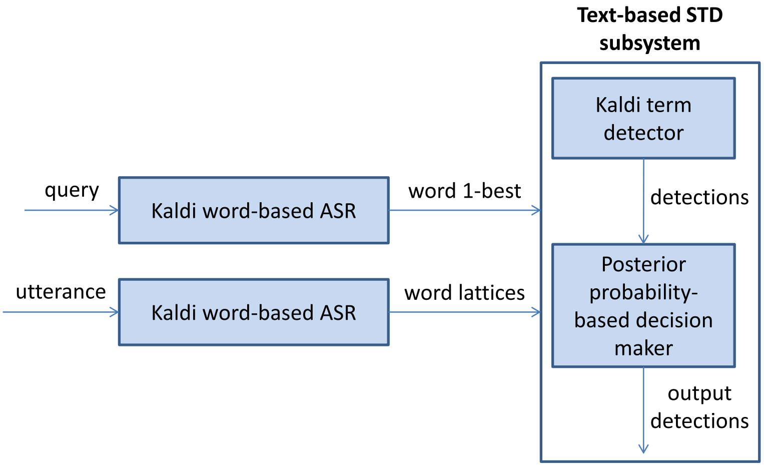

Fig. 2 Kaldi LVCSR-based QbE STD system architecture

The fingerprint extraction comprises two different steps:

- Feature extraction. First, acoustic features are obtained from the audio signal that represents the queries and the utterances. Twelve-dimensional MFCCs augmented with C0 and first and second order derivatives were employed to build a 39-dimensional feature vector. These MFCCs were extracted every 10 ms using a 20-ms sliding window.
- Frame-level fingerprints. The fingerprints corresponding to the acoustic features of each query and utterance frame were obtained from the MFCCs as described in [68]. A convolution mask is used to binarize the acoustic features. Specifically, a mask for finding negative slopes on the spectrogram in two consecutive frames is applied. Given a set of acoustic features $S \in \Re^{I \times J}$, where $S_{i,j}$ is the feature corresponding to energy band i and frame j, the value $F_{i,j}$ of the frame-level fingerprint corresponding to frame j is obtained after applying the convolution mask as follows:

$$F_{i,j} = \begin{cases} 1 & \text{if } \phi > 0 \\ 0 & \text{if } \phi \leq 0 \end{cases} \tag{2}$$

where $\phi = S_{i,j} - S_{i,j+1} + S_{i-1,j} - S_{i-1,j+1}$.

The DTW search was adopted from [29] and comprises four different steps:

- Similarity measure calculation. Euclidean distance between each pair of query frame and utterance frame has been used to compute a matrix that stores the similarity between query and utterance frames. Given a query $Q = \{q_1, \ldots, q_N\}$ and an utterance $S = \{s_1, \ldots, s_M\}$, a similarity matrix $M \in \Re^{N \times M}$ is computed, whose rows and columns correspond to the frames of the query and the utterance, respectively.
- Coarse search. Once the similarity matrix is obtained, a coarse search is carried out to obtain a set of candidate matches. To do so, a sliding window of the size of the query (i.e., N frames) is used with 50 % overlap, and the estimated DTW value of the current window defined by features $(q_{n_i}, \ldots, q_{n_{i+N}})$ is computed as follows:

$$\text{Estimated DTW}(n_i) = \sum_{n=n_i}^{n_{i+N}} \min M_{*,n} \tag{3}$$

where $M_{*,n}$ represents the nth column of the similarity matrix M.
A set of candidates is obtained from this step; specifically, as in [29], the number of candidate detections was selected as the maximum between 100 and the duration of the utterance in seconds. Therefore, those candidate detections that obtain the lowest estimated DTW values are selected.
- Fine search. After selecting the candidate detections in the previous step, an additional DTW search of the query on these candidate positions is conducted. A window of the size of the query is employed for this search, and the actual DTW distance calculated during the DTW search is output.
- Selection of best detections. Once the DTW of the fine search has been carried out, those detections that are separated by at least 0.5 s are kept. Next, the detections whose DTW value is less than a threshold are output as final detections, while the rest are discarded.

Train/dev data were used to train the decision threshold. Next, this threshold was used to hypothesize query detections of the test data.

Phoneme posteriorgram-based QbE STD system

This system is the same as the MFCC fingerprint-based system except that, instead of using MFCC fingerprints for query and utterance representation, phoneme posteriorgrams were employed to build the feature vectors. Specifically, phoneme posteriorgrams [72] were computed for the queries and the utterances by means of a long temporal context-based phoneme recognizer [73]. From all the available phoneme recognizers in [73] (English, Czech, Hungarian, and Russian), the English phoneme recognizer was employed, as it maximized ATWV performance on train/dev data. This phoneme recognizer bases on a TANDEM feature extraction architecture, which merges state phoneme posterior vectors computed from a neural network and PLP coefficients to build the feature vectors. These feature vectors are next fed within a standard GMM/HMM system. Vocal Tract Length Normalization, and speaker based mean and variance normalization are applied in the TANDEM architecture, along with gaussianization and Heteroscedastic Linear Discriminant Analysis for decorrelation and dimensionality reduction. Finally, the same DTW search used in the previous system was employed to hypothesize query detections.

Train/dev data were used to train the decision threshold. Next, this threshold was used to hypothesize query detections of the test data.

System fusion

System fusion combines the output of the three 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 [30]. First, a per-query zero-mean and unit-variance normalization (q-norm) was applied in order to prevent the scores of the individual systems to be in different ranges and to obtain query-independent scores. At this point, fusion is not straightforward, as not all the systems to be fused output a score for every possible trial (i.e., detection); hence, before fusion, the detection problem is transformed into a verification problem. To do so, all the time instants of the detections found by the different systems are first merged and next aligned with the ground-truth to obtain a set of client and impostor trials (i.e., hits and false alarms). As not all the systems may have produced a score for each of these trials, missing scores are hypothesized by computing the average of the other scores for that detection [74]. Once every trial (detection) has a score for each system, fusion is carried out using the Bosaris toolkit [75]; specifically, a logistic regression fusion scheme was trained using the training/development data and then applied to the test data. This procedure results in a new score for each detection, which is then used to output the final detections of the Fusion (SGMM)+Posteriorgram system. The overlapped detections, i.e., detections of different queries at the same time interval, were removed by keeping the query detection with the highest score.

ALBAYZIN QbE STD 2014 evaluation: Fusion (SGMM) system

This system, whose architecture is shown in Fig. 3, consists of the fusion of two different subsystems: the Kaldi LVCSR-based system and the DTW search-based system with an MFCC fingerprint representation of the queries and utterances, both described previously. These two systems are again fused with the fusion strategy described in the previous section.

ALBAYZIN QbE STD 2014 evaluation: Fusion (DNN)+ Posteriorgram system

This system is the same as the Fusion (SGMM)+ Posteriorgram system with the only difference of the type of acoustic models used in the Kaldi LVCSR-based system. In this case, DNN-based acoustic models have been employed, instead of SGMMs. Specifically, a DNN-based context-dependent speech recognizer was trained using the Kaldi toolkit [60] following Karel Vesely's DNN training approach [76]. The network has 6 hidden layers, with 2048 units each. The features employed for DNN training and recognition are computed as follows: 9 frames of 13-dimensional MFCCs are projected down to 40 dimensions using linear discriminant analysis (LDA) and the resulting features are further de-correlated using maximum likelihood linear transform (MLLT); this is followed by speaker normalization using feature-space maximum likelihood linear regression (fMLLR).

ALBAYZIN QbE STD 2014 evaluation: Fusion (DNN) system

This system is the same as the Fusion (SGMM) system with the only difference of the type of acoustic models used in the Kaldi LVCSR-based system. In this case, the same DNN-based LVCSR system employed for the Fusion (DNN)+Posteriorgram system has been used.

ALBAYZIN QbE STD 2014 evaluation: text-based SGMM spoken term detection system (text-based SGMM STD)

A text-based STD system has been used to establish a comparison with the results achieved by the QbE STD systems. This text-based STD system, whose architecture is shown in Fig. 4, follows the same approach as the Kaldi LVCSR-based system described previously in the Fusion (SGMM)+Posteriorgram system, with the only difference being that, in this case, the actual transcription of the query term is given to the system, so the query decoding step is not necessary.

ALBAYZIN QbE STD 2014 evaluation: text-based DNN spoken term detection system (text-based DNN STD)

An additional text-based STD system was built from DNNs. This system, whose architecture is the same as the Text-based SGMM STD system, replaces the SGMM-based acoustic models of the Text-based SGMM STD system by the DNN-based acoustic models described in the Fusion (DNN)+Posteriorgram system.

ALBAYZIN QbE STD 2012 evaluation systems

Four different systems were submitted to the evaluation held in 2012, which were already described in the Iber-SPEECH 2012 proceedings [77]. Here, a brief description of their main characteristics is provided:

System DTW-Zero is based on a DTW zero-resource matching approach. First, Gaussian posteriorgram features are used as signal representation, which are obtained

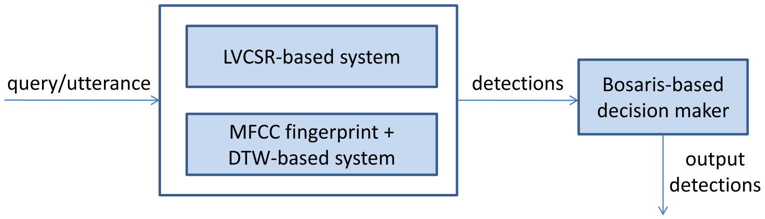
query/utterance
LVCSR-based system
MFCC fingerprint + DTW-based system
detections
Bosaris-based decision maker
output detections

Fig. 3 Fusion (SGMM) system architecture

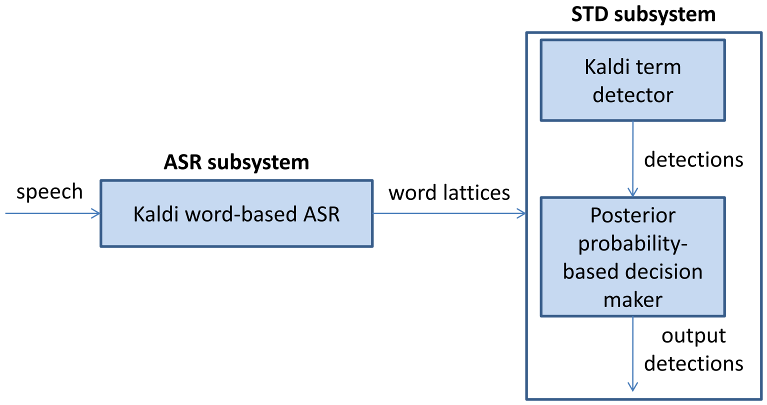

Fig. 4 Text-based SGMM STD system architecture

from a GMM trained from MFCCs. These features are next sent to the subsequence-DTW matching algorithm [19] that employs the cosine distance as similarity measure between query and utterance frames to hypothesize query detections within the utterances.

System P1B-STD is based on an exact match of the phone sequence output by a speech recognizer given the spoken query within the phone lattices corresponding to the utterances. Phone decoders for Czech, Hungarian, and Russian have been employed to produce the phone sequence of each query and the phone lattices. The *Lattice2Multigram* tool [78][9] is used to conduct query search. System P1B-STDa combines the query detections of the Hungarian and Russian phone decoders, and System P1B-STDb combines the query detections of all decoders.

System P1L-STD is based on a search on phone lattices generated from a posteriori phone probabilities. These phone probabilities are obtained by combining the acoustic class probabilities estimated from a GMM-based clustering procedure on the acoustic space and the conditional probabilities of each acoustic class with respect to each phonetic unit [79]. These acoustic classes comprise the set of Spanish phones. Next, the phone probabilities are used in an ASR process to output the phone lattices that represent both the query and the utterance. These query and utterance phone lattices are used for query search, where a search of every path in the query lattice on every path in the utterance lattice is conducted to hypothesize detections. Substitution, insertion, and deletion errors in the lattice matching are allowed. A query-dependent confidence score is assigned to each detection.

System DTW-Spanish employs the same a posteriori phone probabilities as System P1L-STD for query/utterance representation, and these are subsequently used for query search. The query search is based on segmental DTW search [80] with the Kullback-Leibler divergence as similarity measure between query and utterance frames. The same query-dependent confidence score approach used in the P1L-STD system is employed in this system to score each detection.

Results and discussion

ALBAYZIN QbE STD 2014 evaluation

The results of the ALBAYZIN QbE STD 2014 evaluation for train/dev and test data are presented in Tables 9 and 10, respectively. Results show, in general, similar performance for Fusion+Posteriorgram and Fusion systems both for train/dev and test data for SGMM- and DNN-based acoustic models. Paired t tests show that the slight improvement, if any, of the Fusion+Posteriorgram system over the Fusion system is not statistically significant for train/dev and test data ($p \approx 0.3$) for both types of acoustic models. This shows that the phoneme posteriorgram-based system does not pose a complementary behavior

Table 9 Results of the ALBAYZIN QbE STD 2014 evaluation on train/dev data

System ID	MTWV	ATWV	p(FA)	p(Miss)
Fusion (SGMM)+Posteriorgram	0.3023	0.3023	0.00009	0.607
Fusion (SGMM)	0.2957	0.2957	0.00009	0.616
Fusion (DNN)+Posteriorgram	0.3411	0.3388	0.00007	0.590
Fusion (DNN)	0.3394	0.3375	0.00007	0.593
Text-based SGMM STD	0.5639	0.5639	0.00008	0.358
Text-based DNN STD	0.6112	0.6062	0.00006	0.327

Table 10 Results of the ALBAYZIN QbE STD 2014 evaluation on test data

System ID	MTWV	ATWV	p(FA)	p(Miss)
Fusion (SGMM)+Posteriorgram	0.2708	0.2708	0.00006	0.672
Fusion (SGMM)	0.2671	0.2657	0.00005	0.679
Fusion (DNN)+Posteriorgram	0.2894	0.2881	0.00006	0.652
Fusion (DNN)	0.2894	0.2881	0.00006	0.652
Text-based SGMM STD	0.6157	0.6099	0.00006	0.323
Text-based DNN STD	0.6583	0.6469	0.00006	0.282

compared to the MFCC fingerprint-based system, and hence similar results are obtained when one or both are fed to the fusion. In addition, using DNN as acoustic models in the Kaldi LVCSR-based system employed in the fusion outperform the SGMM acoustic models, though the performance gaps are not statistically significant for train/dev data and test data ($p \approx 0.3$). This means that the gains obtained by the DTW-based approach in the fusion compensates the lower performance of the SGMM acoustic models.

Results also show that all the QbE STD systems perform worse than the text-based (SGMM and DNN) STD systems. The improvement of these text-based STD systems over those systems is statistically significant for train/dev data ($p < 10^{-4}$) and for test data ($p < 10^{-6}$). This is mainly due to the use of the correct word transcription of the search query in the text-based STD systems, and indicates that obtaining the word transcription of the query using an ASR system is still problematic. In addition, the improvement obtained by the text-based DNN STD system over the text-based SGMM STD system is statistically significant for train/dev data ($p < 10^{-2}$) and for test data ($p < 0.03$), which shows that the better acoustic modeling plays an important role in the final system performance.

DET curves for train/dev and test data are shown in Figs. 5 and 6, respectively. For train/dev data, it is clear that the Fusion (DNN)+Posteriorgram system performs similar to the Fusion (DNN) system for most of the range, except when false alarm rate is low, where the Fusion (DNN)+Posteriorgram system performs better, and that the DNN-based fused systems outperform the SGMM-based systems counterpart in general. For test data, the Fusion (SGMM)+Posteriorgram and Fusion (DNN)+Posteriorgram systems obtain a slight improvement over the Fusion (SGMM) and Fusion (DNN) systems respectively for most of the range, except for low miss rates, where all systems perform similar. In addition, the Fusion (DNN)+Posteriorgram system performance is worse than that of the Fusion (SGMM)+Posteriorgram when false alarm rate is low, and the same is observed with the Fusion (DNN) and Fusion (SGMM) systems. This confirms our conjecture that the DTW-based approach is able to compensate the lower performance of the SGMM acoustic models.

Foreign query analysis

An analysis on the foreign queries of the test data of the individual and fused systems submitted to the ALBAYZIN QbE STD 2014 evaluation has been conducted, and results are presented in Table 11. There are 5 foreign queries in the test data, with 16 occurrences in total. Due to these low figures, differences in the system results are not, in general, statistically significant for paired t tests. However, we can still shed some light about the performance of the different systems for foreign queries. In general, for QbE STD systems, we observe performance degradation with the LVCSR-based systems compared with the template matching (i.e., DTW)-based systems. MFCC-Fingerprint and Phoneme posteriorgram systems outperform LVCSR SGMM- and LVCSR DNN-based systems. We consider this is due to LVCSR-based systems typically degrade their performance with foreign terms. This is confirmed by the text-based SGMM STD system performance, which is lower than that obtained with the DTW-based systems. However, with the better acoustic modeling of the text-based DNN STD system, the performance is better than that of the DTW-based systems, which confirms the potential of DNNs in STD, especially when the correct term transcription is employed in the search. As expected from the full evaluation results, the fused systems also perform the best for foreign query detection.

ALBAYZIN QbE STD 2012 and 2014 evaluation comparison

Evaluating the systems submitted to the ALBAYZIN QbE STD 2014 evaluation only on the queries of the evaluation held in 2012 produces the results shown in Tables 12 and 13 for train/dev and test data respectively, where these results are compared to those of the systems submitted to the 2012 evaluation.

For train/dev data, all the fusion-based systems improve the rest of the QbE STD systems. Paired t tests show that this improvement is statistically significant ($p < 10^{-3}$). For test data, the same behavior is observed. These results constitute a relevant progress for QbE STD in Spanish from the evaluation held in 2012. The best performance of these fusion-based systems is mainly due to two reasons: (1) The use of a robust LVCSR system for Spanish language in terms of acoustic model (SGMM and DNN) and language model (trained from a large variety of text sources), and (2) all the systems employ a fusion of different kinds of systems. Since word transcription-based systems and template matching-based systems present a complementary behavior for QbE STD, the combination of both yields improvements for QbE STD. When comparing Fusion (DNN)+Posteriorgram, Fusion (DNN), Fusion (SGMM)+Posteriorgram, and Fusion (SGMM) systems,

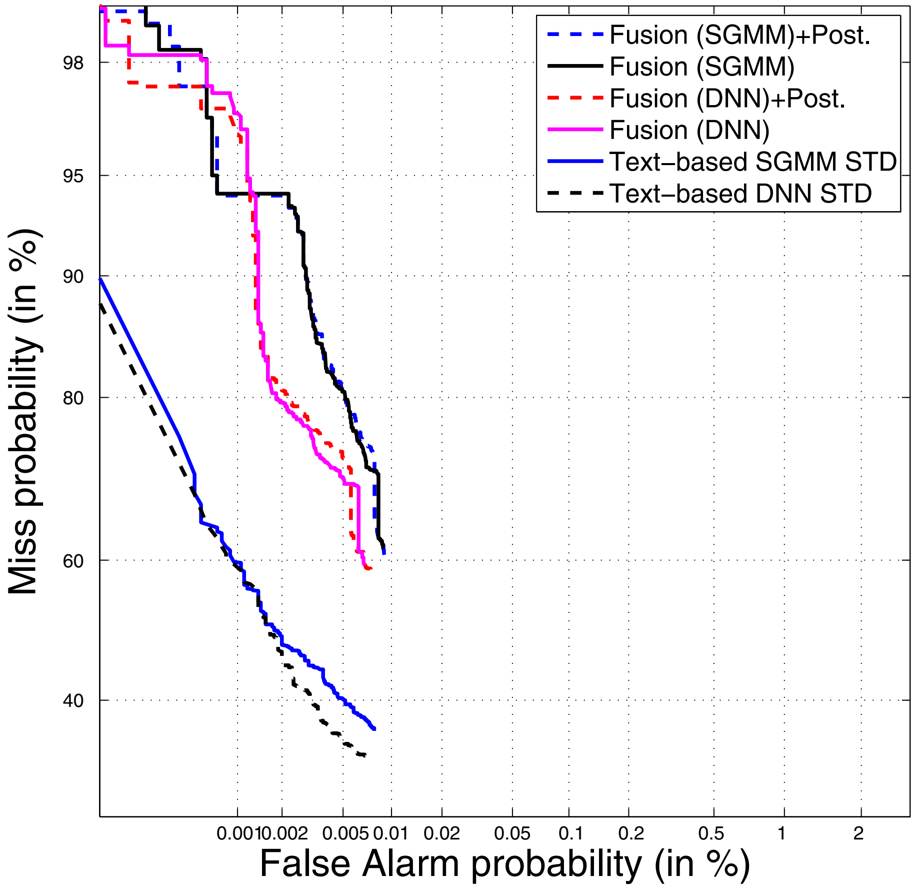

Fig. 5 DET curves of the systems submitted to the ALBAYZIN QbE STD 2014 evaluation for train/dev data. *Post.* posteriorgram

we observe that the DNN-based systems outperform the SGMM-based systems, though there is no significant difference between them ($p \approx 0.4$ for train/dev data and $p \approx 0.7$ for test data), which is consistent with the results presented in the 2014 evaluation.

Among the systems that do not employ system fusion (all the fused systems integrate an LVCSR system), the DTW-Spanish system obtains the best overall performance for train/dev and test data. The query-dependent score normalization produces the smallest difference between MTWV and ATWV. This indicates that the threshold is well set. Paired t tests show that this best performance of the DTW-Spanish system is statistically significant for train/dev data ($p < 10^{-6}$) and test data ($p < 10^{-3}$) compared to the other single systems. This system is language-dependent, and this is probably one of the reasons for the best performance of the DTW-Spanish system. In addition, the similarity measure used to conduct the segmental DTW search (Kullback-Leibler divergence) fits very well the posterior probabilities computed in the feature extraction stage. Aiming at building a language-independent QbE STD system, the DTW-Zero system deserves special mention, since it obtains the best performance in terms of MTWV on test data. In this case, a better threshold setting is needed to get nearer ATWV to MTWV.

The corresponding DET curves for train/dev and test data for the systems submitted to the evaluations held in 2012 and 2014 are shown in Figs. 7 and 8, respectively. We observe similar trends for both sets of data. Fusion (DNN)+Posteriorgram, Fusion (SGMM)+Posteriorgram, Fusion (DNN), and Fusion (SGMM) systems perform the best for almost all the range, except when the false alarm rate is low, where DTW-Zero system performs the best. For test data, the Fusion (SGMM)+Posteriorgram system performs better than the Fusion (SGMM) system for almost all the range, and the Fusion (DNN)+Posteriorgram system outperforms the Fusion (DNN) system, which is consistent with the results obtained in the 2014 evaluation.

Analyzing the DET curves for the single systems, it is clear that the DTW-Zero system, which employs a language-independent approach for QbE STD, outperforms the rest of the systems for train/dev and test

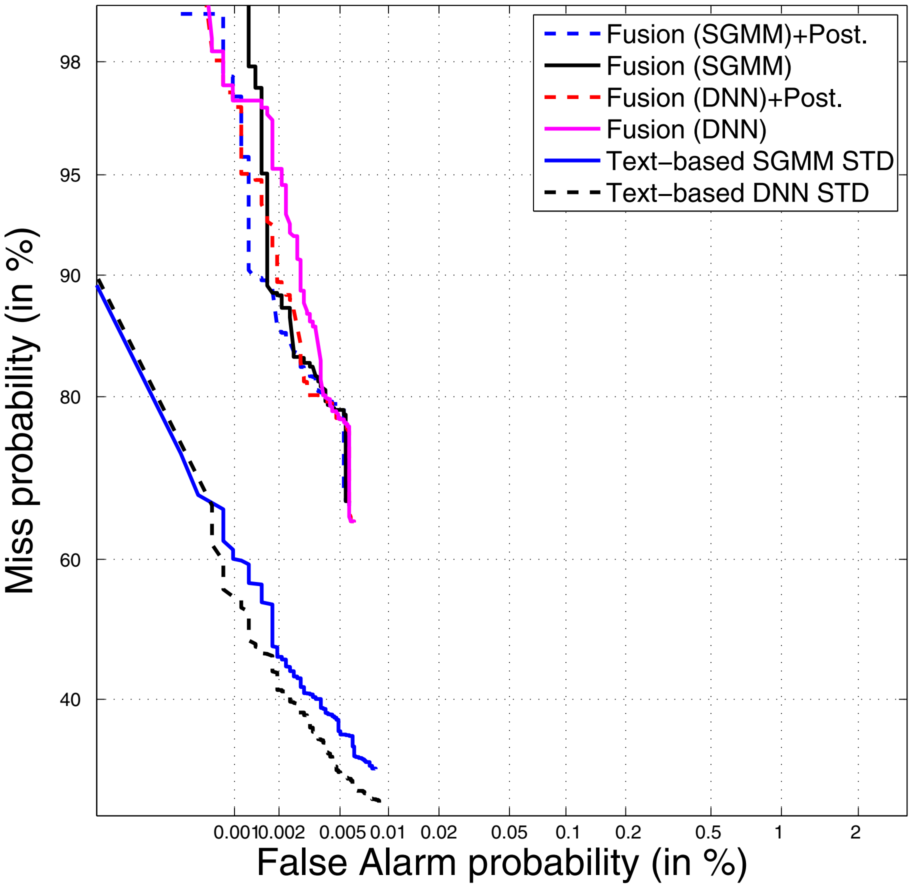

Fig. 6 DET curves of the systems submitted to the ALBAYZIN QbE STD 2014 evaluation for test data. *Post.* posteriorgram

data, except at the best operating point of the DTW-Spanish system, where this performs better than the DTW-Zero system. This makes the DTW-Zero system interesting to face the language independency issue in QbE STD.

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

System ID	MTWV	ATWV	*p*(FA)	*p*(Miss)
Fusion (SGMM)+Posteriorgram	0.5167	0.5167	0	0.483
Fusion (SGMM)	0.5167	0.5167	0	0.483
Fusion (DNN)+Posteriorgram	0.5500	0.5500	0	0.450
Fusion (DNN)	0.5500	0.5500	0	0.450
LVCSR-SGMM	0.2667	0.2667	0	0.733
LVCSR-DNN	0.3000	0.3000	0	0.700
MFCC-Fingerprint	0.3833	0.3833	0	0.617
Phoneme Posteriorgram	0.3833	0.3833	0	0.617
Text-based SGMM STD	0.3675	0.3400	0.00008	0.550
Text-based DNN STD	0.4225	0.4225	0.00003	0.500

Toward a language-independent STD system

From the systems submitted to the ALBAYZIN QbE STD 2012 and 2014 evaluations, insights about the feasibility of building a language-independent STD system can be gained. By comparing the best language-independent

Table 12 Results of the ALBAYZIN QbE STD 2012 and 2014 evaluations on the ALBAYZIN QbE STD 2012 train/dev data

System ID	MTWV	ATWV	*p*(FA)	*p*(Miss)
Fusion (SGMM)+Posteriorgram	0.2850	0.2850	0.00011	0.610
Fusion (SGMM)	0.2824	0.2803	0.00010	0.619
Fusion (DNN)+Posteriorgram	0.3572	0.3514	0.00007	0.578
Fusion (DNN)	0.3579	0.3571	0.00006	0.580
DTW-Zero	0.0455	0.0455	0.00002	0.930
P1B-STDa	0.0128	0.0128	0.00000	0.986
P1B-STDb	0.0092	0.0092	0.00000	0.990
P1L-STD	0.0000	0.0000	0.00000	1.000
DTW-Spanish	0.0612	0.0612	0.00005	0.893
Text-based SGMM STD	0.6866	0.6866	0.00009	0.226
Text-based DNN STD	0.7440	0.7398	0.00006	0.192

Table 13 Results of the ALBAYZIN QbE STD 2012 and 2014 evaluations on the ALBAYZIN QbE STD 2012 test data

System ID	MTWV	ATWV	*p*(FA)	*p*(Miss)
Fusion (SGMM)+Posteriorgram	0.2691	0.2691	0.00006	0.676
Fusion (SGMM)	0.2691	0.2691	0.00006	0.676
Fusion (DNN)+Posteriorgram	0.2815	0.2815	0.00007	0.647
Fusion (DNN)	0.2815	0.2815	0.00007	0.647
DTW-Zero	0.0436	0.0122	0.00000	0.952
P1B-STDa	0.0055	0.0031	0.00001	0.983
P1B-STDb	0.0075	0.0047	0.00000	0.990
P1L-STD	0.0000	-0.0678	0.00000	1.000
DTW-Spanish	0.0238	0.0217	0.00009	0.884
Text-based SGMM STD	0.6795	0.6627	0.00006	0.256
Text-based DNN STD	0.7299	0.7148	0.00007	0.199

QbE STD system (DTW-Zero) with the text-based DNN STD system, we can claim that building a language-independent STD system with a performance similar to that of a language-dependent STD system is still far from being achieved. This means that more research is needed in QbE STD to approximate language-independent to language-dependent STD systems in highly difficult speech domains such as spontaneous speech.

Challenge of the QbE STD task

From the results obtained by all the systems submitted to the ALBAYZIN QbE STD 2012 and 2014 evaluations, we can claim that building a QbE STD system with a performance near to that of a text-based STD system for Spanish is still difficult. The ATWV performance obtained by the best QbE STD system (ATWV= 0.2894) compared to that of the best text-based STD system (ATWV= 0.6583) confirms this. There are still many issues that must be solved in the future, depending on the type of the system.

On the one hand, for systems based on LVCSR, an accurate word transcription of the query must be obtained. Otherwise, the QbE STD system performance dramatically drops, as we have seen in the 2014 evaluation results.

On the other hand, for systems that rely on template matching, a robust template that efficiently represents the queries and the utterances is necessary. In addition, a reliable search algorithm that hypothesizes query detections is also necessary to output as many hits as

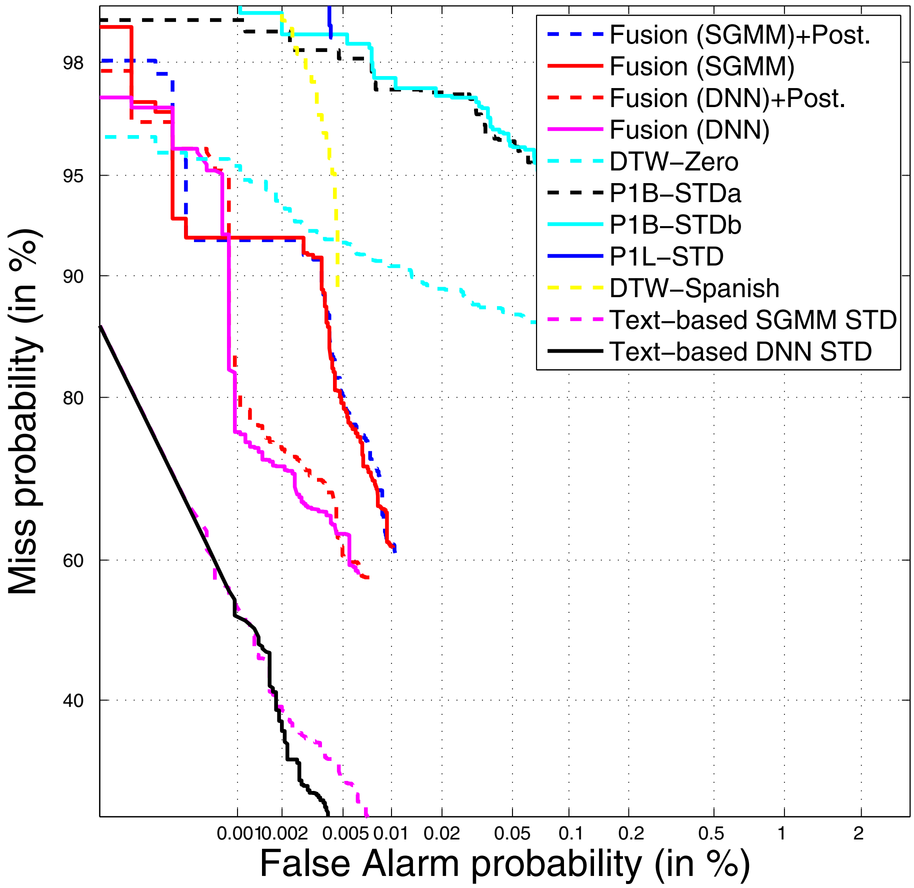

Fig. 7 DET curves of the systems submitted to the ALBAYZIN QbE STD 2012 and 2014 evaluations on train/dev data. *Post.* posteriorgram

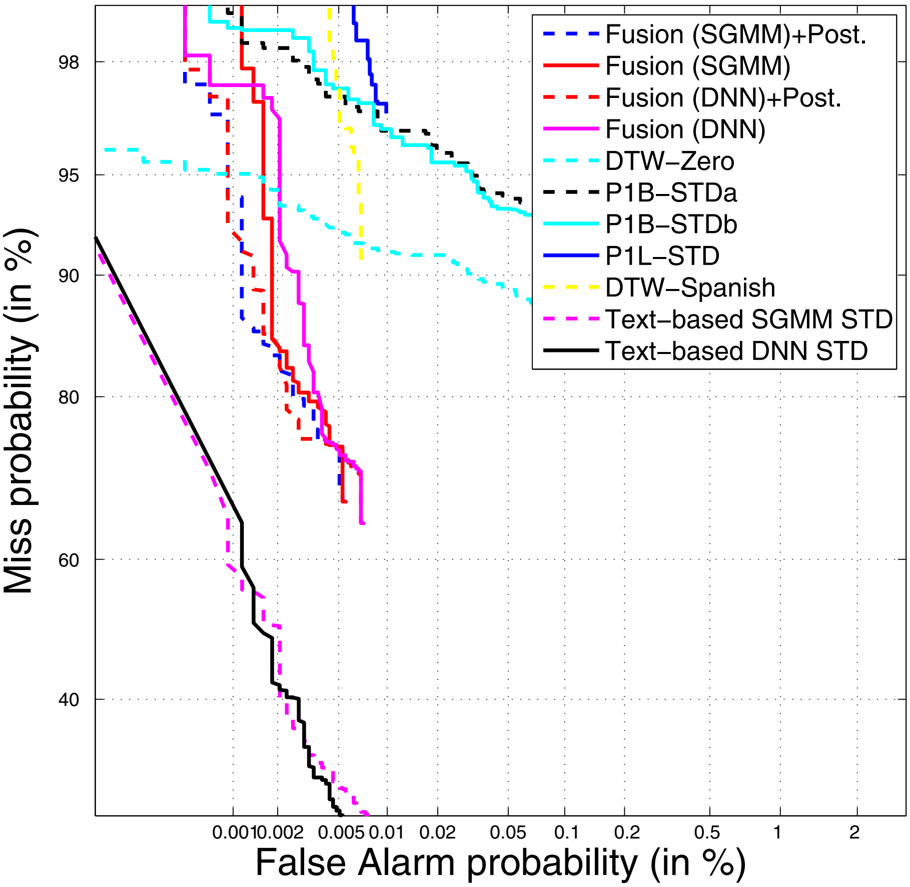

Fig. 8 DET curves of the systems submitted to the ALBAYZIN QbE STD 2012 and 2014 evaluations on test data. *Post.* posteriorgram

possible while maintaining a reasonably low number of FAs.

All the systems must also deal with the type of speech of the evaluation, mainly spontaneous speech, which represents an important challenge for the QbE STD task. In this way, as in standard ASR systems, special attention must be paid to phenomena such as disfluences, hesitations, and noises. The QbE STD system performance can possibly be enhanced by including some pre-processing steps that deal with these phenomena.

Lessons learned

The different types of systems submitted to the first and second QbE STD evaluations in Spanish provide significant lessons that should be taken into account for forthcoming evaluation editions.

The first ALBAYZIN QbE STD evaluation held in 2012 received systems that are mainly language-independent, whereas the evaluation held in 2014 focused on language-dependent systems. Results presented in this paper have shown performance differences between language-dependent and language-independent QbE STD systems. Moreover, for the language-dependent systems, results have also shown performance differences between systems that employ fusion of word transcription- and template matching-based systems, and systems that do not. Therefore, organizers will thoroughly think about dividing the ALBAYZIN QbE STD evaluation for future editions into two different subtasks. The first subtask is suitable for systems that are language-dependent, whereas the second one will be for systems that are language-independent, aiming at building a language-independent STD system and evaluating it in Spanish.

The second ALBAYZIN QbE STD evaluation incorporates multi-word and foreign queries to the evaluation. However, there were just a few multi-word and foreign queries in this second evaluation. Future evaluations should include more multi-word and foreign queries.

To compare the system results of the ALBAYZIN QbE STD evaluations held in 2012 and 2014, the same evaluation metric (ATWV) has been employed. However, in the recent MediaEval 2014 QbE Search on Speech evaluation [39], a different metric called normalized cross entropy cost (C_{nxe}) has been employed. This metric requires calibrated likelihood ratios, and hence, participants will be allowed to submit calibrated likelihood ratios for future evaluation editions. Moreover, the NTCIR-11 QbE STD evaluation also employed a different evaluation metric (F-measure). This metric, contrary to ATWV, assigns the same cost to precision and recall values and hence allows comparing the systems from another perspective. Therefore, organizers will probably propose to use both evaluation metrics (C_{nxe} and F-measure) as secondary metrics for next evaluation editions.

The first and second ALBAYZIN QbE STD evaluations focused on searching a train/dev query list in train/dev speech data and searching a test query list in test speech data. In future evaluations, the cross-data query search should also be considered. The purpose of this cross-data search is to see how critical tuning is for the different systems. For example, searching test queries in train/dev speech data could be enhanced by unsupervised adaptation, whereas searching train/dev queries in test speech data can measure the generalization capability of the systems on unseen data with the same query list for which good classifiers could have been developed.

Regarding the data preparation, organizers used the database of MAVIR project consisting of recordings of seminars and round tables organized at the general meetings of the project. 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 ALBAYZIN QbE STD evaluation held in 2012, organizers focused on single-word queries in Spanish, but in the second edition, organizers added multi-word and foreign queries 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 query 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. Organizers are currently preparing more data in order to perform a new and more challenging evaluation in 2016. Besides using new data, organizers will probably reuse the same MAVIR data to assess technology improvements on a comparable basis.

Conclusions

This paper presented the systems submitted to the ALBAYZIN QbE STD 2014 evaluation along with two systems that conduct text-based STD and compared these with the systems submitted to the ALBAYZIN QbE STD 2012 evaluation. Four different Spanish research groups (TID, GTTS, ELiRF, and GTM) took part in the evaluations. Different kinds of systems were submitted to the evaluations: fusion-based systems (Fusion (SGMM)+Posteriorgram, Fusion (DNM)+Posteriorgram, Fusion (SGMM), and Fusion (DNN)), template matching-based systems (DTW-Zero and DTW-Spanish), and subword transcription-based systems (P1B-STD and P1L-STD).

Results show that the best performance is obtained from fused systems that combine word transcription- and template matching-based systems. These fused systems employ the target language (i.e., Spanish) information.

From the results presented in the ALBAYZIN QbE STD 2014 evaluation, we can claim that significant performance gains have been obtained compared to the ALBAYZIN QbE STD 2012 evaluation. This confirms the progress on QbE STD technology for Spanish language. However, when comparing the QbE STD results with the results of the text-based DNN STD system presented in this paper, it is clear that there is still ample room for improvement to approximate the performance of QbE STD to that of text-based STD. This encourages organizers to maintain this evaluation in the next ALBAYZIN evaluation campaign.

Endnotes

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

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

[3] http://iberspeech2012.ii.uam.es/.

[4] http://iberspeech2014.ulpgc.es/.

[5] MAVIR was a project funded by the Madrid region that coordinated several research groups and companies working on information retrieval (http://www.mavir.net)

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

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

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

[9] http://homepages.inf.ed.ac.uk/v1dwang2/public/tools/index.html.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Systems submitted to the second QbE STD evaluation for Spanish language are presented. They are hybrid, language-dependent systems. Systems are compared with the systems submitted to the first QbE STD evaluation for Spanish language (which were mainly single, language-independent systems), and show significant improvements. Lessons learned from both evaluations are also presented, which aim to highlight interesting clues for evaluation organizers.

Acknowledgements

This research was funded by the Spanish Government ('SpeechTech4All Project' TEC2012-38939-C03-01 and 'CMC-V2 Project' TEC2012-37585-C02-01), the Galician Government through the research contract GRC2014/024 (Modalidade: Grupos de Referencia Competitiva 2014) and 'AtlantTIC Project' CN2012/160, and also by the Spanish Government and the European Regional Development Fund (ERDF) under project TACTICA.

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.

Received: 2 September 2015 Accepted: 4 January 2016
Published online: 13 January 2016

References

1. T Zhang, C-CJ Kuo, in *Hierarchical classification of audio data for archiving and retrieving*. Proc. of ICASSP (IEEE, Washington DC, USA, 1999), pp. 3001–3004
2. M Helén, T Virtanen, in *Query by example of audio signals using Euclidean distance between Gaussian Mixture Models*. Proc. of ICASSP (IEEE, Washington DC, USA, 2007), pp. 225–228
3. M Helén, T Virtanen, Audio query by example using similarity measures between probability density functions of features. EURASIP, Journal on Audio, Speech, and Music Processing. **2010**, 2–1212 (2010)
4. G Tzanetakis, A Ermolinskyi, P Cook, in *Pitch histograms in audio and symbolic music information retrieval*. Proc. of ISMIR (ISMIR, Paris, France, 2002), pp. 31–38
5. W-H Tsai, H-M Wang, in *A query-by-example framework to retrieve music documents by singer*. Proc. of ICME (IEEE, Washington DC, USA, 2004), pp. 1863–1866
6. TK Chia, KC Sim, H Li, HT Ng, in *A lattice-based approach to query-by-example spoken document retrieval*. Proc. of ACM SIGIR (ACM, New York, USA, 2008), pp. 363–370
7. A Muscariello, G Gravier, F Bimbot, in *Zero-resource audio-only spoken term detection based on a combination of template matching techniques*. Proc. of Interspeech (ISCA, Baixas, France, 2011), pp. 921–924
8. J Tejedor, M Fapšo, I Szöke, Černocký, F Grézl, Comparison of methods for language-dependent and language-independent query-by-example spoken term detection. ACM Trans. Inf. Syst. **30**(3), 18–11834 (2012)
9. G Mantena, X Anguera, in *Speed improvements to information retrieval-based dynamic time warping using hierarchical k-means clustering*. Proc. of ICASSP (IEEE, Washington DC, USA, 2013), pp. 8515–8519
10. G Mantena, S Achanta, K Prahallad, Query-by-example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(5), 946–955 (2014)
11. J Vavrek, M Pleva, M Lojka, P Viszlay, Kiktová, D Hládek, J Juhár, in *TUKE at MediaEval 2013 spoken web search task*. Proc. of MediaEval (CEUR, Aachen, Germany, 2013), pp. 73–1732
12. R Jarina, M Kuba, R Gubka, M Chmulik, M Paralic, in *UNIZA system for the spoken web search task at MediaEval 2013*. Proc. of MediaEval (CEUR, Aachen, Germany, 2013), pp. 79–1792
13. A Ali, MA Clements, in *Spoken web search using and ergodic hidden Markov model of speech*. Proc. of MediaEval (CEUR, Aachen, Germany, 2013), pp. 86–1862
14. A Buzo, H Cucu, C Burileanu, in *SpeeD@MediaEval 2014: Spoken term detection with robust multilingual phone recognition*. Proc. of MediaEval (CEUR, Aachen, Germany, 2014), pp. 72–1722
15. S Kesiraju, G Mantena, K Prahallad, in *IIIT-H system for MediaEval 2014 QUESST*. Proc. of MediaEval (CEUR, Aachen, Germany, 2014). pp. 76–1762
16. J Takahashi, T Hashimoto, R Konno, S Sugawara, K Ouchi, S Oshima, T Akyu, Y Itoh, in *An IWAPU STD system for OOV query terms and spoken queries*. Proc. of NTCIR-11 (National Institute of Informatics, Tokyo, Japan, 2014), pp. 384–389
17. M Makino, A Kai, in *Combining subword and state-level dissimilarity measures for improved spoken term detection in NTCIR-11 SpokenQuery&Doc task*. Proc. of NTCIR-11 (National Institute of Informatics, Tokyo, Japan, 2014), pp. 413–418
18. M Gubian, L Boves, M Versteegh, in *Calibration of distance measures for unsupervised query-by-example*. Proc. of Interspeech (ISCA, Baixas, France, 2013), pp. 2639–2643
19. X Anguera, M Ferrarons, in *Memory efficient subsequence DTW for query-by-example spoken term detection*. Proc. of ICME (IEEE, Washington DC, USA, 2013)
20. H Wang, T Lee, in *The CUHK spoken web search system for MediaEval 2013*. Proc. of MediaEval (CEUR, Aachen, Germany, 2013), pp. 68–1682
21. M Bouallegue, G Senay, M Morchid, D Matrouf, G Linares, R Dufour, in *LIA@MediaEval 2013 spoken web search task: An I-Vector based approach*. Proc. of MediaEval (CEUR, Aachen, Germany, 2013), pp. 77–1772
22. LJ Rodriguez-Fuentes, A Varona, M Penagarikano, G Bordel, M Diez, in *GTTS systems for the SWS task at MediaEval 2013*. Proc. of MediaEval (CEUR, Aachen, Germany, 2013), pp. 83–1832
23. H Wang, T Lee, C-C Leung, B Ma, H Li, in *Using parallel tokenizers with DTW matrix combination for low-resource spoken term detection*. Proc. of ICASSP (IEEE, Washington DC, USA, 2013), pp. 8545–8549
24. H Wang, T Lee, in *CUHK system for QUESST task of MediaEval 2014*. Proc. of MediaEval (CEUR, Aachen, Germany, 2014), pp. 73–1732
25. J Proenca, A Veiga, F Perdigão, in *The SPL-IT query by example search on speech system for MediaEval 2014*. Proc. of MediaEval (CEUR, Aachen, Germany, 2014), pp. 74–1742
26. P Yang, C-C Leung, L Xie, B Ma, H Li, in *Intrinsic spectral analysis based on temporal context features for query-by-example spoken term detection*. Proc. of Interspeech (ISCA, Baixas, France, 2014), pp. 1722–1726
27. B George, A Saxena, G Mantena, K Prahallad, B Yegnanarayana, in *Unsupervised query-by-example spoken term detection using bag of acoustic words and non-segmental dynamic time warping*. Proc. of Interspeech (ISCA, Baixas, France, 2014), pp. 1742–1746
28. TJ Hazen, W Shen, CM White, in *Query-by-example spoken term detection using phonetic posteriorgram templates*. Proc. of ASRU (IEEE, Washington DC, USA, 2009), pp. 421–426
29. A Abad, RF Astudillo, I Trancoso, in *The L2F spoken web search system for MediaEval 2013*. Proc. of MediaEval (CEUR, Aachen, Germany, 2013), pp. 85–1852
30. A Abad, LJ Rodríguez-Fuentes, M Penagarikano, A Varona, G Bordel, in *On the calibration and fusion of heterogeneous spoken term detection systems*. Proc. of Interspeech (ISCA, Baixas, France, 2013), pp. 20–24
31. I Szöke, M Skácel, L Burget, in *BUT QUESST 2014 system description*. Proc. of MediaEval (CEUR, Aachen, Germany, 2014), pp. 62–1622
32. P Yang, H Xu, X Xiao, L Xie, C-C Leung, H Chen, J Yu, H Lv, L Wang, SJ Leow, B Ma, ES Chng, H Li, in *The NNI query-by-example system for MediaEval 2014*. Proc. of MediaEval (CEUR, Aachen, Germany, 2014), pp. 69–1692
33. I Szöke, L Burget, F Grézl, JH Černocký, L Ondel, in *Calibration and fusion of query-by-example systems - BUT SWS 2013*. Proc. of ICASSP (IEEE, Washington DC, USA, 2014), pp. 7849–7853
34. H Wang, T Lee, C-C Leung, B Ma, H Li, Acoustic segment modeling with spectral clustering methods. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(2), 264–277 (2015)
35. C-T Chung, W-N Hsu, C-Y Lee, L-S Lee, in *Enhancing automatically discovered multi-level acoustic patterns considering context consistency with applications in spoken term detection*. Proc. of ICASSP (IEEE, Washington DC, USA, 2015), pp. 5231–5235
36. NIST, The Ninth Text REtrieval Conference (TREC 9) (2000). http://trec.nist.gov. Accessed 8 January 2016
37. H Joho, K Kishida, in *Overview of the NTCIR-11, SpokenQuery&Doc Task*. Proc. of NTCIR-11 (National Institute of Informatics, Tokyo, Japan, 2014), pp. 1–7
38. X Anguera, F Metze, A Buzo, I Szöke, LJ Rodriguez-Fuentes, in *The spoken web search task*. Proc. of MediaEval (CEUR, Aachen, Germany, 2013), pp. 1–2
39. X Anguera, LJ Rodriguez-Fuentes, I Szöke, A Buzo, F Metze, in *Query by example search on speech at Mediaeval 2014*. Proc. of MediaEval (CEUR, Aachen, Germany, 2014), pp. 1–2
40. 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. Accessed 8 January 2016
41. B Taras, C Nadeu, Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion. EURASIP Journal on Audio, Speech, and Music Processing. **2011**, 1–1110 (2011)
42. M Zelenák, H Schulz, J Hernando, Speaker diarization of broadcast news in Albayzin 2010 evaluation campaign. EURASIP Journal on Audio, Speech, and Music Processing. **2012**, 19–1199 (2012)

43. LJ Rodríguez-Fuentes, M Penagarikano, A Varona, M Díez, G Bordel, in *The Albayzin 2010 Language Recognition Evaluation*. Proc. of Interspeech (ISCA, Baixas, France, 2011), pp. 1529–1532
44. 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, Journal on Audio, Speech, and Music Processing. **2013**, 23–12317 (2013)
45. F Méndez, L Docío, M Arza, F Campillo, in *The Albayzin 2010 text-to-speech evaluation*. Proc. of FALA (Spanish Thematic Network on Speech Technology, Madrid, Spain, 2010), pp. 317–340
46. J Billa, KW Ma, JW McDonough, G Zavaliagkos, DR Miller, KN Ross, A El-Jaroudi, in *Multilingual speech recognition: the 1996 Byblos callhome system*. Proc. of Eurospeech (ISCA, Baixas, France, 1997)
47. H Cuayahuitl, B Serridge, in *Out-of-vocabulary word modeling and rejection for spanish keyword spotting systems*. Proc. of MICAI (Springer, London, United Kingdom, 2002), pp. 156–165
48. M Killer, S Stuker, T Schultz, in *Grapheme based speech recognition*. Proc. of Eurospeech (ISCA, Baixas, France, 2003), pp. 3141–3144
49. J Tejedor, *Contributions to Keyword Spotting and Spoken Term Detection For Information Retrieval in Audio Mining. PhD thesis, Universidad Autónoma de Madrid, Madrid, Spain*. (Universidad AutÃ^{3}noma de Madrid, Madrid, Spain, 2009)
50. L Burget, P Schwarz, M Agarwal, P Akyazi, K Feng, A Ghoshal, O Glembek, N Goel, M Karafiat, D Povey, A Rastrow, RC Rose, S Thomas, in *Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models*. Proc. of ICASSP (IEEE, Washington DC, USA, 2010), pp. 4334–4337
51. 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)
52. J Li, X Wang, B Xu, in *An empirical study of multilingual and low-resource spoken term detection using deep neural networks*. Proc. of Interspeech (ISCA, Baixas, France, 2014), pp. 1747–1751
53. 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. Accessed 8 January 2016
54. JG Fiscus, J Ajot, JS Garofolo, G Doddingtion, in *Results of the 2006 spoken term detection evaluation*. Proc. of SSCS (ACM, New York, USA, 2007), pp. 45–50
55. A Martin, G Doddington, T Kamm, M Ordowski, M Przybocki, in *The DET curve in assessment of detection task performance*. Proc. of Eurospeech (ISCA, Baixas, France, 1997), pp. 1895–1898
56. 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. Accessed 8 January 2016
57. N Rajput, F Metze, in *Spoken web search*. Proc. of MediaEval (CEUR, Aachen, Germany, 2011), pp. 1–2
58. F Metze, E Barnard, M Davel, Heerden C van, X Anguera, G Gravier, N Rajput, in *The spoken web search task*. Proc. of MediaEval (CEUR, Aachen, Germany, 2012), pp. 1–2
59. NTCIR-11 Spoken Query and Spoken Document Retrieval Task Organizers, Definition of SQ-STD Task at NTCIR-11 SpokenQuery&Doc (2014). http://www.nlp.cs.tut.ac.jp/~sdpwg/ntcir11/SQ-STD.pdf. Accessed 8 January 2016
60. 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 *The Kaldi speech recognition toolkit*. Proc. of ASRU (IEEE, Washington DC, USA, 2011)
61. L Docío-Fernández, A Cardenal-López, C García-Mateo, in *TC-STAR 2006 automatic speech recognition evaluation: The uvigo system*. Proc. of TC-STAR Workshop on Speech-to-Speech Translation (META-NET, Berlin, Germany, 2006)
62. A Stolcke, in *SRILM - an extensible language modeling toolkit*. Proc. of ICSLP (ISCA, Baixas, France, 2002), pp. 901–904
63. D Povey, M Hannemann, G Boulianne, L Burget, A Ghoshal, M Janda, M Karafiat, S Kombrink, P Motlicek, Y Qian, K Riedhammer, K Vesely, NT Vu, in *Proc. of ICASSP*. Generating exact lattices in the WFST framework (IEEE, Washington DC, USA, 2012), pp. 4213–4216
64. G Chen, S Khudanpur, D Povey, J Trmal, D Yarowsky, O Yilmaz, in *Quantifying the value of pronunciation lexicons for keyword search in low resource languages*. Proc. of ICASSP (IEEE, Washington DC, USA, 2013), pp. 8560–8564
65. VT Pham, NF Chen, S Sivadas, H Xu, I-F Chen, C Ni, ES Chng, H Li, in *System and keyword dependent fusion for spoken term detection*. Proc. of SLT (IEEE, Washington DC, USA, 2014), pp. 430–435
66. D Can, M Saraclar, Lattice indexing for spoken term detection. IEEE Trans. Audio Speech Lang. Process. **19**(8), 2338–2347 (2011)
67. P Lopez-Otero, L Docio-Fernandez, C Garcia-Mateo, in *Introducing a framework for the evaluation of music detection tools*. Proc. of LREC (European Language Resources Association, Paris, France, 2014), pp. 568–572
68. C Neves, A Veiga, Sá, F Perdigão, in *Audio fingerprinting system for broadcast streams*. Proc. of ConfTele, vol. 1, (Santa Maria da Feira, Instituto de Telecomunicações, Campus Universitãrio de Santiago, Aveiro, Portugal, 2009), pp. 481–484
69. K Seyerlehner, G Widmer, T Pohle, M Sched, in *Proc. of the 10th Conference on Digital Audio Effects*. Automatic music detection in television productions (LaBRI, UniversitÃ© Bordeaux, Bordeaux, France, 2007)
70. S Kim, E Unal, S Narayanan, in *Music fingerprint extraction for classical music cover song identification*. Proc. of ICME (IEEE, Washington DC, USA, 2008), pp. 1261–1264
71. J Haitsma, T Kalker, in *A highly robust audio fingerprinting system*. Proc. of ISMIR (ISMIR, Paris, France, 2002), pp. 107–115
72. TJ Hazen, W Shen, CM White, in *Query-by-example spoken term detection using phonetic posteriorgram templates*. Proc. of ASRU (IEEE, Washington DC, USA, 2009), pp. 421–426
73. P Schwarz, *Phoneme recognition based on long temporal context. PhD thesis, Brno University of Technology*. (Brno University of Technology, Brno, Czech Republic, 2009)
74. A Abad, RF Astudillo, in *The L2F spoken web search system for mediaeval 2012*. Proc. of MediaEval (CEUR, Aachen, Germany, 2012), pp. 9–10
75. N Brümmer, E de Villiers, The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing. Technical report (2011). https://sites.google.com/site/nikobrummer. Accessed 8 January 2016
76. Veselý, A Ghoshal, L Burget, D Povey, in *Sequence-discriminative training of deep neural networks*. Proc. of Interspeech (ISCA, Baixas, France, 2013), pp. 2345–2349
77. IberSPEECH 2012, "VII Jornadas en Tecnología del Habla" and "III Iberian SLTech Workshop" (2012). http://iberspeech2012.ii.uam.es. Accessed 8 January 2016
78. D Wang, S King, J Frankel, Stochastic pronunciation modelling for out-of-vocabulary spoken term detection. IEEE Trans. Audio Speech Lang. Process. **19**(4), 688–698 (2011)
79. JA Gómez, E Sanchis, MJ Castro-Bleda, in *Automatic speech segmentation based on acoustical clustering*. Proc. of the Joint IAPR International Conference on Structural, Syntactic, and Statistical Pattern Recognition (Springer, London, United Kingdom, 2010), pp. 540–548
80. A Park, JR Glass, in *Towards unsupervised pattern discovery in speech*. Proc. of ASRU (IEEE, Washington DC, USA, 2005), pp. 53–58