Tejedor and Toledano
EURASIP Journal on Audio, Speech, and Music Processing (2024) 2024:15
https://doi.org/10.1186/s13636-024-00334-w

EURASIP Journal on Audio, Speech, and Music Processing

EMPIRICAL RESEARCH

Open Access

Whisper-based spoken term detection systems for search on speech ALBAYZIN evaluation challenge

Javier Tejedor[1*] and Doroteo T. Toledano[2]

Abstract

The vast amount of information stored in audio repositories makes necessary the development of efficient and automatic methods to search on audio content. In that direction, search on speech (SoS) has received much attention in the last decades. To motivate the development of automatic systems, ALBAYZIN evaluations include a search on speech challenge since 2012. This challenge releases several databases that cover different acoustic domains (i.e., spontaneous speech from TV shows, conference talks, parliament sessions, to name a few) aiming to build automatic systems that retrieve a set of terms from those databases. This paper presents a baseline system based on the Whisper automatic speech recognizer for the spoken term detection task in the search on speech challenge held in 2022 within the ALBAYZIN evaluations. This baseline system will be released with this publication and will be given to participants in the upcoming SoS ALBAYZIN evaluation in 2024. Additionally, several analyses based on some term properties (i.e., in-language and foreign terms, and single-word and multi-word terms) are carried out to show the Whisper capability at retrieving terms that convey specific properties. Although the results obtained for some databases are far from being perfect (e.g., for broadcast news domain), this Whisper-based approach has obtained the best results on the challenge databases so far so that it presents a strong baseline system for the upcoming challenge, encouraging participants to improve it.

Keywords Search on speech, Spoken term detection, Whisper, ALBAYZIN evaluations

1 Introduction

The huge amount of information stored in audio repositories makes necessary to build efficient methods to retrieve it. In this direction, search on speech (SoS) has been considered an upmost area within which the required technology can be effectively constructed. Therefore, this area has been receiving much interest from decades from spoken document retrieval (SDR) [1–13], keyword spotting (KWS) [14–27], and query-by-example spoken term detection (QbE-STD) [28–42] tasks. Within search on speech, spoken term detection (STD) has also emerged as a powerful task that aims to retrieve speech data from a textual query (henceforth term). Due to the enormous potential of the STD task, this has also been receiving much attention for years from different companies and research groups such as IBM [43–48], BBN [49–51], SRI and OGI [52–54], BUT [55–57], Microsoft [58, 59], QUT [60, 61], JHU [62–65], Fraunhofer IAIS/NTNU/TUD [66], NTU [67, 68], IDIAP [69], among others [70–73].

Spoken term detection systems are mainly composed of two different subsystems, as it is shown in Fig. 1. First, an offline subsystem so-called automatic speech recognition (ASR) subsystem provides the transcription

*Correspondence:
Javier Tejedor
javier.tejedornoguerales@ceu.es
[1] Institute of Technology, Universidad San Pablo-CEU, CEU Universities, Urbanización Montepríncipe, Boadilla del Monte 28668, Spain
[2] AUDIAS, Electronics and Communication Technology Department, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Av. Francisco Tomás y Valiente, 11, Madrid 28049, Spain

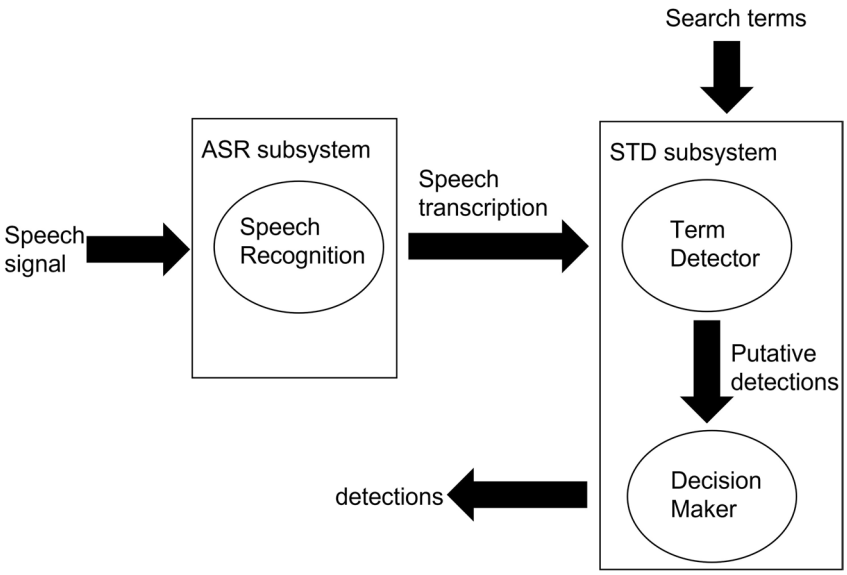

Fig. 1 Architecture of an STD system

according to the speech content. Second, an on-line subsystem so-called STD subsystem integrates a term detector to hypothesize detections from the ASR output and a decision maker that ascertains detections based on confidence scoring.

1.1 Spoken term detection evaluations around the world

Worldwide evaluations have been shown to be a powerful tool to promote research in different disciplines. The National Institute of Standards and Technology (NIST) launched the first STD evaluation in 2006 focusing on English, Arabic, and Mandarin languages and different acoustic domains such as telephone speech, broadcast news speech, and meeting room speech [74]. The IARPA BABEL program, NIST Open Keyword Search (OpenKWS) evaluation series [75–78], and Open Speech Analytics Technologies (OpenSAT) evaluation series have also allowed to compare STD technology within a common and standard framework on noisy conditions and under-resourced languages such as Cantonese, Pashto, Tagalog, Swahili, Tamil, to name a few [46, 48, 50, 51, 65, 79–93] during the 2010s.

On the other hand, STD evaluations have also been held in the framework of the NTCIR conferences for Japanese language and spontaneous speech from workshops between 2011 and 2016 [94–97].

Recently, in 2021, Interspeech conference held the Auto-KWS 2021 Challenge [98], in which participants had to build a customized keyword spotting system where the target device can only be awakened by an enrolled speaker with his/her specified keyword. Therefore, this challenge combined keyword spotting and speaker verification tasks in a single evaluation. English and Chinese languages were addressed in this challenge.

Focusing on search on speech and Spanish language, SoS ALBAYZIN evaluation challenges were launched from 2012 every 2 years aiming to promote research on search on speech in Spanish. One of the tasks that this evaluation challenge addresses is STD. Participants, therefore, have a common framework to evaluate their STD systems on a single but complete evaluation that integrates several acoustic domains and term sets [99–102]. This ALBAYZIN STD task differs from the other evaluations held around the world in the target language and the different domains that the ALBAYZIN STD task provides to participants to measure the generalization capability of the submitted systems in different domains. In addition, each new challenge adds more complexity with respect to the previous one by including new and more challenging terms, new databases, etc. To allow for comparison among different evaluation editions, one of the databases (the MAVIR database [102] described in Section 3.1) has been included in all the editions of the SoS ALBAYZIN evaluation challenges.

1.2 Motivation and organization of this paper

Whisper ASR system from OpenAI company has been shown to be an outstanding speech recognizer for many languages spoken around the world [103]. Aiming to provide a strong baseline for the STD task in

the upcoming SoS ALBAYZIN evaluation challenge to be held during IberSPEECH conference in 2024, this paper presents an STD system based on the Whisper ASR system and evaluates it on the data provided by the ALBAYZIN evaluation challenge held in 2022. To the best of our knowledge, this is first time that the Whisper ASR system is used for the STD task and is evaluated on real data from a certain evaluation campaign on Spanish language.

The rest of the paper is organized as follows: Section 2 presents an overview of the systems that have been submitted to the STD task within the SoS ALBAYZIN evaluation challenges. The databases used for the challenge held in 2022 are presented in Section 3, along with the term selection procedure. The evaluation metrics employed to measure system performance are presented in Section 4. The Whisper ASR and STD systems are presented in Section 5, and the results of the ASR and STD systems are discussed in Section 6. Additional analyses based on term properties are presented in Section 7, and the paper is concluded in Section 8.

2 Overview of the systems submitted to the STD task within the search on speech ALBAYZIN challenge

A wide variety of STD systems were submitted to the STD task within the SoS ALBAYZIN evaluation challenges from 2012. Most of them are largely based on the hidden Markov model toolkit (HTK) [99] and Kaldi toolkit for ASR decoding [99–102]. Additionally, participants also submitted systems based on their own speech recognition system [99].

Most of the systems employ word-based ASR to obtain word lattices or 1-best word output within which the terms are detected. However, there are also systems that address the out-of-vocabulary (OOV) issue of those word-based systems by resorting to sub-word (i.e., phoneme) ASR [99, 100].

Additionally, fusion techniques were also shown to be effective for system performance improvement [99–101]. Synthesis-based approaches to convert textual terms into acoustic queries and search based on QbE-STD were also researched in [100, 101].

3 Databases

Three different databases that convey different domains have been used for the spoken term detection task within the search on speech ALBAYZIN 2022 evaluation challenge: MAVIR database, which consists of spontaneous speech from conference talks, RTVE22 database, which consists of speech from several TV programs within the broadcast news domain, and SPARL22 database, which consists of speech from several Spanish parliament sessions.

Table 1 Development and test term list characteristics for MAVIR database

Term list	Development	Test
#INL (occ.)	354 (959)	208 (2071)
#OOL (occ.)	20 (55)	15 (50)
#SINGLE (occ.)	340 (984)	198 (2093)
#MULTI (occ.)	34 (30)	25 (28)

occ. number of occurrences (in brackets), *INL* in-language, *OOL* out-of-language, *SINGLE* single-word terms, *MULTI* multi-word terms. The term length in the development term list varies between 5 and 27 graphemes (single-word term length varies between 5 and 16 graphemes, and multi-word term length varies between 7 and 27 graphemes). The term length in the test term list varies between 4 and 28 graphemes (single-word term length varies between 4 and 16 graphemes, and multi-word term length varies between 7 and 28 graphemes)

3.1 MAVIR database

This database comprises data from Spanish conference talks within MAVIR workshops held in 2006, 2007, and 2008 [104]. Specifically, three different datasets were released for the challenge purposes: training dataset, which amounts to about 4 h of speech, development dataset, which amounts to 1 h of speech, and test dataset, with 2 h of speech, in total. The number of terms and the number of occurrences of each term along with other meaningful database information are presented in Table 1 for development and test data. It must be noted that neither a list of terms nor their timestamps in the corresponding audio have been provided for the training data.

More information regarding this database can be found in [102], which is freely available for research purposes.[1]

3.2 RTVE22 database

This database comprises data from Spanish TV programs recorded from 1960 to the present[2]. Specifically, three different datasets were released for the challenge purposes: training dataset, which amounts to about 900 h of speech, development dataset, which amounts to 15 h of speech, and test dataset, with 5 h of speech, in total. The number of terms and the number of occurrences of each term along with other meaningful database information are presented in Table 2 for development and test data. It must be noted that neither a list of terms nor their timestamps in the corresponding audio have been provided for the training data.

More information of this database can be found in [105].

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

[2] https://catedrartve.unizar.es/rtvedatabase.html

Table 2 Development and test term list characteristics for RTVE22 database

Term list	Development	Test
#INL (occ.)	307 (1151)	188 (930)
#OOL (occ.)	91 (351)	72 (109)
#SINGLE (occ.)	380 (1280)	217 (985)
#MULTI (occ.)	18 (222)	43 (54)

occ. number of occurrences (in brackets), *INL* in-language, *OOL* out-of-language, *SINGLE* single-word terms, *MULTI* multi-word terms. The term length in the development term list varies between 4 and 25 graphemes (single-word term length varies between 4 and 20 graphemes, and multi-word term length varies between 7 and 25 graphemes). The term length in the test term list varies between 3 and 28 graphemes (single-word term length varies between 3 and 12 graphemes, and multi-word term length varies between 8 and 28 graphemes)

Table 3 Test term list characteristics for SPARL22 database

Term list	Test
#INL (occ.)	262 (1557)
#OOL (occ.)	20 (46)
#SINGLE (occ.)	258 (1560)
#MULTI (occ.)	24 (43)

occ. number of occurrences (in brackets)INL in-language, *OOL* out-of-language, *SINGLE* single-word terms, *MULTI* multi-word terms. The term length in the test term list varies between 2 and 26 graphemes (single-word term length varies between 2 and 17 graphemes, and multi-word term length varies between 5 and 26 graphemes)

3.3 SPARL22 database

This database comprises spontaneous speech recorded from Spanish parliament sessions held from 2016[3]. Only test data, which amount to 2 h of speech, comprise the single evaluation dataset, since this database is used in the challenge to evaluate the system performance in an *unseen* domain. The number of terms and the number of occurrences of each term along with other meaningful database information are presented in Table 3 for these test data.

More information of this database can be found in [102].

3.4 Term list selection procedure

The term list selection procedure plays an important role within search on speech evaluations, since it should carefully take into account different search scenarios. To do so, the terms chosen for search were manually selected from each speech database individually to include high-occurrence terms, low-occurrence terms, in-language (INL) (i.e., Spanish) terms, out-of-language (OOL) (i.e., foreign) terms, single-word and multi-word terms, and terms of different lengths for all the databases. Regarding the OOL terms, those cover English language, since they comprise the highest OOL coverage across the different databases. For MAVIR development data, those OOL terms represent a 0.52% of the total number of word occurrences (i.e., all the words spoken on that dataset). For MAVIR test data, the OOL terms represent a 0.25% of the total number of word occurrences. For RTVE22 development data, there is a 2.76% of OOL occurrences with respect to the total number of word occurrences, and for RTVE22 test data, there is a 0.24% of OOL occurrences with respect to all the words spoken on those data. For SPARL22 test data, the coverage of the OOL occurrences is of 0.25% with respect to the total number of word occurrences. In every database, a term may not have any occurrence or appear one or more times in the speech data.

[3] https://www.congreso.es/es/archivo-audiovisual

Table 4 Overall term list characteristics

Term list	Terms	Occurrences
#INL	1319	6668
#OOL	218	611
#SINGLE	1393	6902
#MULTI	144	377

INL in-language, *OOL* out-of-language, *SINGLE* single-word terms, *MULTI* multi-word terms

Table 4 summarizes the number of in-language, out-of-language, single-word, and multi-word terms along with their number of occurrences of all the databases, in which it can be seen that OOL coverage reaches a 9.2% of the INL coverage.

Table 5 collects the number of terms that are shared between the different databases and datasets.

4 Evaluation metrics

In STD systems, an occurrence output by a certain system is called a *detection*; if the detection corresponds to an actual occurrence, it is called a *hit*; otherwise, it is called a *false alarm*. If an actual occurrence is not detected, this is called a *miss*. The actual term weighted value (ATWV) metric proposed by NIST [74] has been used as the main metric for system evaluation. This metric integrates both the hit rate and false alarm rate of each term and averages over all the terms, as shown in Eq. 1:

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

where Δ denotes the set of terms and $|\Delta|$ is the number of terms in the corresponding speech dataset. N_{hit}^T and N_{FA}^T denote the numbers of hits and false alarms of the term T respectively, and N_{true}^T represents the number of actual

Table 5 Number of terms that are shared between databases and datasets

Dataset	MAVIR dev.	MAVIR test	RTVE22 dev.	RTVE22 test	SPARL22
MAVIR dev.	–	80	18	10	18
MAVIR test	80	–	3	5	12
RTVE22 dev.	18	3	–	25	15
RTVE22 test	10	5	25	–	8
SPARL22 test	18	12	15	8	–

dev. development

occurrences of term T in the audio. L denotes the audio length in seconds, and β is a weight factor set to 999.9, as in the ATWV proposed by NIST [49]. This weight factor causes an emphasis placed on recall compared to precision with a ratio 10:1.

For system evaluation, a detection will be labeled as correct in case this appears within ± 15-s interval from the ground-truth timestamp. This interval is higher than that proposed originally by NIST [49] to encourage participants to build end-to-end systems, for which timestamp estimation was not so precise by the evaluation designing time.

The ATWV is computed with the actual decision threshold established by the system, which is usually tuned on development data. Sub-optimal threshold setting may imply a loss in performance that may hinder the capabilities of the systems. For that reason, an additional metric, called maximum term weighted value (MTWV) [74], has also been used to evaluate the upper-bound system performance regardless the decision threshold. The MTWV, therefore, is computed using the optimal threshold for the given dataset and the confidence scores output by the system.

Additionally, p(Miss) and p(FA) values, which represent the probability of miss and FA of the system as defined in Eqs. 2 and 3, respectively, are also reported for system evaluation.

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

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

where N_{hit} is the number of hits obtained by the system, N_{true} is the actual number of occurrences of all the terms in the audio, N_{FA} is the number of FAs in the system, and L denotes the audio length (in seconds).

In addition to ATWV, MTWV, p(Miss), and p(FA) figures, NIST also proposed a detection error tradeoff (DET) curve [106] that evaluates the performance of a system at various miss/FA ratios. They are also presented in this paper for system comparison.

For computing the MTWV, ATWV, p(Miss), and p(FA) figures, along with the DET curves, the NIST STD evaluation tool [107] was employed.

5 Whisper-based ASR and spoken term detection systems

The systems presented in this manuscript for STD are largely based on the Whisper ASR, with the suitable modifications to address STD. This section first describes the Whisper-based ASR system and then presents the Whisper-timestamped modification and the STD systems based on that.

5.1 Whisper ASR system

Whisper is an ASR system presented by OpenAI in September 2021[4] and released in open source along with several pre-trained models[5]. It is intentionally based on a standard encoder-decoder transformer structure, as shown in Fig. 2.

The transformer architecture was introduced by Vaswani et al. in 2017 [108]. The key element of the transformer is the multi-head attention block. It is composed of a number of attention heads working in parallel. Each attention head takes as input a query (Q) and a sequence of key-value (K,V) pairs and provides as output a weighted average of the values, where the weights (more precisely, the attention weights) are computed according to a compatibility function (which is essentially a dot product) between the query and the keys. In this way, the output of each attention head is a view of the whole input, focusing on different parts of the input according to the attention weights computed from the query. In the encoder part, keys, values, and queries are the same (keys and values are the whole input from the previous layer, and the query for each time is the input at that time), as can be seen in Fig. 2. This type of

[4] https://openai.com/research/whisper

[5] https://github.com/openai/whisper

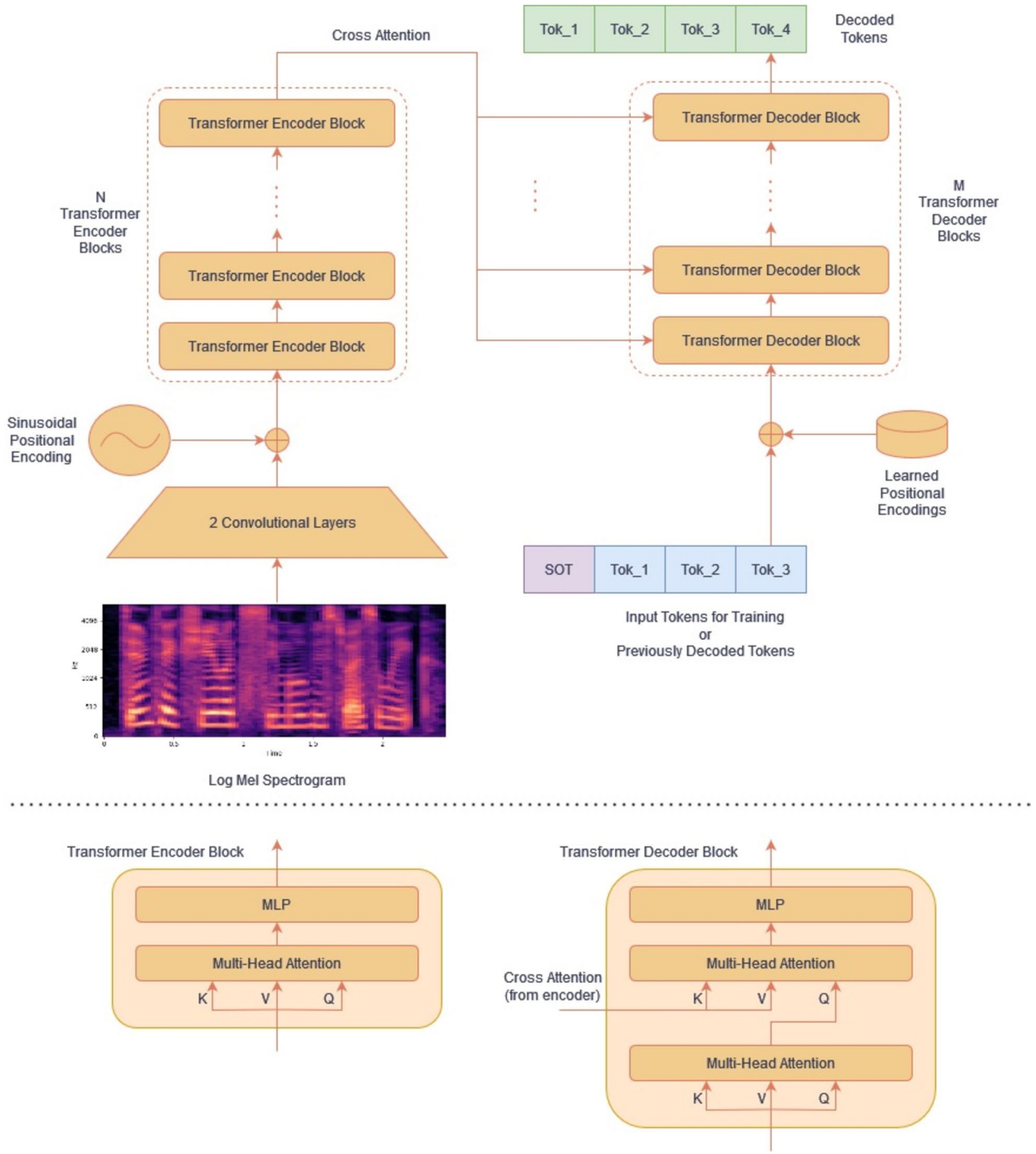

Fig. 2 Whisper encoder-decoder transformer structure. Top: Overall structure. Bottom: Detail of the encoder and decoder blocks. MLP, multi-layer perceptron; K, keys; V, values; Q, queries [103]

attention is called *self-attention* because the attention weights are computed using only the input. In the decoder part (see Fig. 2), there are multi-head attention layers operating with *self-attention*, but there are also multi-head attention layers computing *cross-attention* weights, in which the keys and the values are the output of the encoder, while the query comes from the previous layer of the decoder and essentially provides information about the sequence decoded so far. In this way, the *cross-attention* weights indicate which parts of the input

are important to predict the next token. In the case of automatic speech recognition, these *cross-attention* weights give approximate information of the time span of the input corresponding to each decoded token.

The main innovation of Whisper with respect to previous open-source ASR systems is the type and amount of training data: Whisper has been trained on 680,000 h of weakly supervised speech collected from the web and has been trained in a multi-task (covering ASR, language identification and ASR+translation tasks) and multi-language way (covering almost one hundred languages including English, Chinese, Spanish, French, German, Italian, Polish, Arabic, Finnish, Mongolian, Tamil, Thai, Urdu, to name a few). Whisper models can identify the language, transcribe the speech, and recognize and translate many languages into English. OpenAI also provides small models with less computational requirements (see Table 1 in [103] for more details). In our case, we have employed the medium-size model for all the experiments. One of the most relevant results of Whisper is that is has been shown to be competitive with the most advanced ASR systems without any task adaptation, just using the provided models in a *zero-shot* fashion. Besides, its performance in English has been shown to be comparable to that of the most advanced commercial ASR systems and even similar to professional human transcribers. As other end-to-end neural ASR systems, Whisper uses byte pair encoding (BPE) [109, 110] to codify text into tokens that can be single characters, fragments of words, or even whole words or sequences of words. These tokens are defined in a data-driven way from a text database and a target number of tokens (typically between 32K and 64K). BPE is an effective way to interpolate between word-based lexicons and character-based ASR systems. It can model frequent words as single tokens and less frequent words as sequences of word fragments or even sequences of individual characters. A consequence of particular interest for STD is that Whisper can produce any sequence of characters as output. This contrasts with traditional ASR systems that used (necessarily limited) word-based lexicons to generate the predictions and could not output words that were not present in their vocabulary, giving rise to the problem of OOV words that could never been recognized (and could not be directly found in STD). This problem is completely avoided with Whisper. All these features make Whisper an excellent candidate for a strong ASR baseline. More information about the Whisper ASR system can be found in [103].

5.2 Whisper-based STD systems

Due to the power of Whisper on ASR tasks, STD systems built on top of that can be effectively used for search on speech tasks. Although Whisper can predict timestamps, the timestamps provided by Whisper correspond to the captions on which the system has been trained (i.e., it provides start and ending times of groups of words or sentences). For STD it is necessary to have, at least, the timestamps of all the transcribed words and the Whisper models released by OpenAI do not provide that information directly. Fortunately, a tool addressing this issue called Whisper-timestamped[6] has been released very recently. Whisper-timestamped [111] is largely based on the Whisper ASR system [103], but it includes additional mechanisms to provide word timestamps as well as word-based confidences, both of which are essential for STD. Word timestamps implementation is based on the cross-attention weights computed by the decoder in the Whisper ASR system (see Fig. 2), which, as explained in the previous section, contain valuable information about the time alignment of the predicted words. This information is processed with a set of heuristics along with the dynamic time warping (DTW) algorithm [112] to find the proper alignment between the audio and the transcription. The confidence score for each word is estimated from the average log probability-based method of the Whisper ASR (i.e., after taking the log softmax of the network's output logits, the average log probability of the tokens chosen by the decoding is used as a confidence score).

Two different systems based on Whisper-timestamped have been built, which aim to provide a baseline for the upcoming SoS ALBAYZIN evaluation and STD task. The first system aims to detect terms that are composed by a single word (single-word STD system), whereas the second system (multi-word STD system) also addresses multi-word term detection. Both systems are described next.

5.2.1 Single-word STD system (Whisper (single))

The *Whisper (single)* STD system, whose system architecture is shown in Fig. 3, is a single-word term detection approach based on the Whisper-timestamped tool. First, the Whisper-timestamped tool is run as an ASR subsystem to obtain both the words recognized in the speech signal, along with their timestamps and confidence scores. An excerpt example in *json* format of the Whisper-timestamped tool output that shows the words recognized in a certain time interval along with the words, timestamps, and scores is shown next:

[6] https://github.com/linto-ai/whisper-timestamped

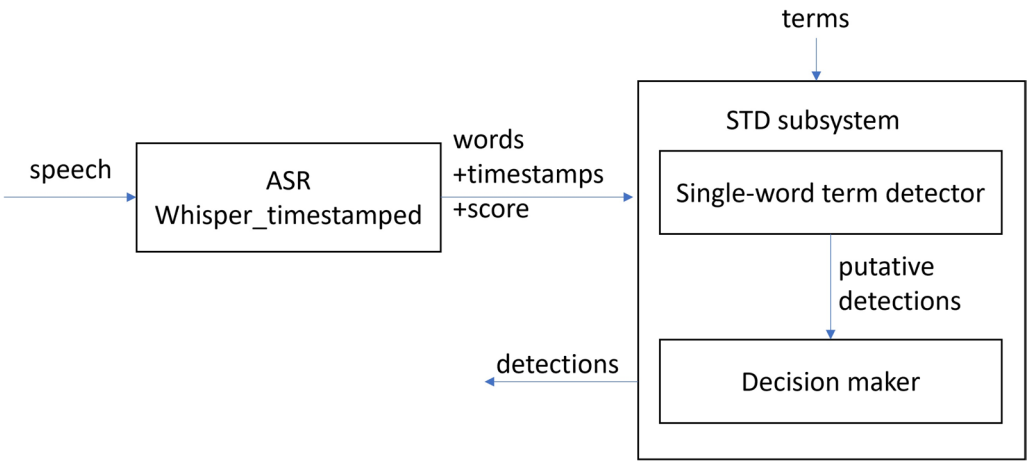

Fig. 3 Architecture of the Whisper (single) STD system

"segments": [
{
"id": 0,
"seek": 0,
"start": 0.34,
"end": 7.16,
"text": "Muchas gracias Julio, buenos días. Adolfo Corujo, León, Dilbert, para qué le diría que improvisara.",
"tokens": [
35669,
16611,
7174,
1004,
11,
...
],
"temperature": 0.0,
"avg_logprob": -0.1973336197947728,
"compression_ratio": 1.482394366197183,
"no_speech_prob": 0.09470756351947784,
"confidence": 0.78,
"words": [
{
"text": "Muchas",
"start": 0.34,
"end": 0.64,
"confidence": 0.755
},
{
"text": "gracias",
"start": 0.64,
"end": 1.02,
"confidence": 0.988
},
{
"text": "Julio,",
"start": 1.02,
"end": 1.68,
"confidence": 0.645
},
{
"text": "buenos",
"start": 1.98,
"end": 2.24,
"confidence": 0.967
},
{
"text": "días.",
"start": 2.24,
"end": 2.68,
"confidence": 0.669
},
{
"text": "Adolfo",
"start": 3.16,
"end": 3.6,
"confidence": 0.812
},
{
"text": "Corujo,",
"start": 3.6,
"end": 3.96,
"confidence": 0.958
},
{
"text": "León,",
"start": 4.1,
"end": 4.46,
"confidence": 0.884
},
{
"text": "Dilbert,",

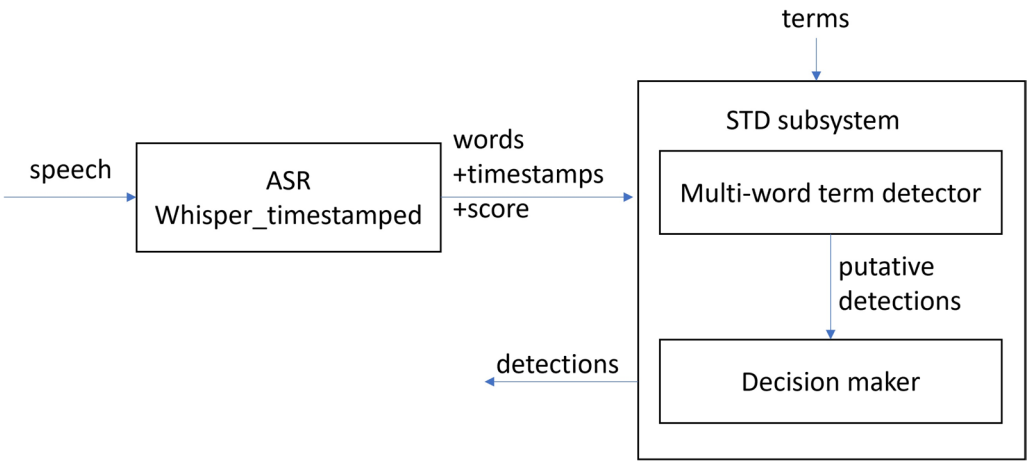

Fig. 4 Architecture of the Whisper (multi) STD system

"start": 4.64,
"end": 5.18,
"confidence": 0.692
},
{
"text": "para",
"start": 5.84,
"end': 5.98,
"confidence": 0.487
},
...

Then, a single-word term detector searches for the word of the given term in the output of the ASR subsystem. To do so, the *json* output format file corresponding to the transcription of the given speech file is converted into a new *detection* file that contains the list of putative detections along with their timestamps and scores. The excerpt example of this *detection* file is shown next:

gracias 0.64 0.38 0.988 mavir03
dilbert 4.64 0.54 0.692 mavir03,

where the first column refers to the term, the second column refers to the start timestamp, the third column refers to the duration of the detection (which is obtained from the *start* and *end* values of the *json* file), the fourth column refers to the confidence score, and the last column refers to the audio file of the detection.

Finally, a decision maker ascertains detections by assigning YES decision to all the occurrences output by the term detection block (i.e., all the occurrences in the *detection* file). The output of this decision maker is an 'stdlist' XML-format file, according to the NIST STD evaluation definition [113]. The excerpt of the example corresponding to the *detection* file is shown next:

<stdlist termlist_filename="resultMAVIR.xml" indexing_time="1.000" language="spanish" index_size="1" system_id="STD">
<detected_termlist termid="gracias" term_search_time="24.3" oov_term_count="1">
<term file="mavir03" channel="1" tbeg="0.64" dur="0.38" score="0.988" decision="YES"/>
</detected_termlist>
<detected_termlist termid="dilbert" term_search_time="24.3" oov_term_count="1">
<term file="mavir03" channel="1" tbeg="4.64" dur="0.54" score="0.692" decision="YES"/>
</detected_termlist>

It must be noted that the term detector of this system takes the words output by the ASR subsystem "as is"; therefore, it is not able to retrieve any multi-word term, since the output of the ASR subsystem consists of single words.

ASR experiments were run according to the following command:

whisper_timestamped audiofilename -o outputdirectory –model medium –language Spanish −−accurate,

where *audiofilename* is the audio file to recognize, *outputdirectory* is the directory where the recognition output is stored, *medium* refers to the type of model employed in the decoding (medium model in our case for system simplicity), *Spanish* is the target language of the audio files, and *−−accurate* refers to the default configuration parameters when running ASR decoding (best_of=5, beam_search=5, temperature_increment_on_fallback=0.2).

5.2.2 Multi-word STD system (Whisper (multi))

The *Whisper (multi)* STD system, whose architecture is shown in Fig. 4, is built on top of the *Whisper (single)* STD system to address multi-word term detection.

Both *Whisper (single)* and *Whisper (multi)* STD systems share the same blocks, except the term detector. Therefore, to construct the multi-word term detector,

the words output by the Whisper-timestamped tool in the *Whisper (single)* system are considered so that a multi-word term is detected (therefore a new detection appears in the STD system) in case all the words that compose the term are correctly recognized according to the word order in the given term. Given the following excerpt of the *json* format output of the Whisper_timestamped tool:

"segments": [
{
"id": 0,
"seek": 0,
"start": 0.34,
"end": 7.16,
"text": "Muchas gracias Julio, buenos días. Adolfo Corujo, León, Dilbert, para qué le diría que improvisara.",
"tokens": [
35669,
16611,
7174,
1004,
11,
...
],
"temperature": 0.0,
"avg_logprob": -0.1973336197947728,
"compression_ratio": 1.482394366197183,
"no_speech_prob": 0.09470756351947784,
"confidence": 0.78,
"words": [
{
"text": "Muchas",
"start": 0.34,
"end": 0.64,
"confidence": 0.755
},
{
"text": "gracias",
"start": 0.64,
"end": 1.02,
"confidence": 0.988
},
{
"text": "Julio,",
"start": 1.02,
"end": 1.68,
"confidence": 0.645
},
{
"text": "buenos",
"start": 1.98,
"end": 2.24,
"confidence": 0.967
},
{
"text": "días.",
"start": 2.24,
"end": 2.68,
"confidence": 0.669
},
{
"text": "Adolfo",
"start": 3.16,
"end": 3.6,
"confidence": 0.812
},
{
"text": "Corujo,",
"start": 3.6,
"end": 3.96,
"confidence": 0.958
},
{
"text": "León,",
"start": 4.1,
"end": 4.46,
"confidence": 0.884
},
{
"text": "Dilbert,",
"start": 4.64,
"end": 5.18,
"confidence": 0.692
},
{
"text": "para",
"start": 5.84,
"end': 5.98,
"confidence": 0.487
},
...,

the corresponding excerpt example of the multi-word term detector is show next:

gracias 0.64 0.38 0.988 mavir03
adolfo_corujo 3.16 0.8 0.885 mavir03
dilbert 4.64 0.54 0.692 mavir03,

where the term *adolfo corujo*, which consists of two words (that are separated with _ symbol) and therefore is missed in the *Whisper (single)* STD system, is detected with this enhanced multi-word term detector.

As it can be seen in the multi-word term detector output, the start and end timestamps given to the multi-word term detection are the initial timestamp of the first word of the term and the end timestamp of the last word of the term. Regarding the confidence score given to the detection, this is computed as the average

Table 6 ASR results

	MAVIR-dev	MAVIR-test	RTVE22-dev	RTVE22-test	SPARL22-test
WER	8.99%	9.95%	11.48%	12.26%	5.76%

MAVIR-dev MAVIR development data, *MAVIR-test* MAVIR test data, *RTVE22-dev* RTVE22 development data, *RTVE22-test* RTVE22 test data, *SPARL22-test* SPARL22 test data, *WER* word error rate

of all the individual confidence scores for every word in the term.

The decision maker assigns YES decision to all the detections (as in the *Whisper (single)* STD system), so that the corresponding output excerpt of that block is as follows:

<stdlist termlist_filename="resultMAVIR.xml" indexing_time="1.000" language="spanish" index_size="1" system_id="STD">

<detected_termlist termid="gracias" term_search_time="24.3" oov_term_count="1">

<term file="mavir03" channel="1" tbeg="0.64" dur="0.38" score="0.988" decision="YES"/>

</detected_termlist>

<detected_termlist termid="adolfo_corujo" term_search_time="24.3" oov_term_count="1">

<term file="mavir03" channel="1" tbeg="3.16" dur="0.8" score="0.885" decision="YES"/>

</detected_termlist>

<detected_termlist termid="dilbert" term_search_time="24.3" oov_term_count="1">

<term file="mavir03" channel="1" tbeg="4.64" dur="0.54" score="0.692" decision="YES"/>

</detected_termlist>.

6 Results and discussion

This section presents the results obtained with both the Whisper-based ASR system and the two STD systems built on top of the ASR system. It must be noted that the development file within the RTVE22 data named *millennium-20171211.aac* was removed from the experiments, due to the high amount of errors included in the ground-truth, which made impossible a fair ASR and STD evaluation. In addition, the file named *LN24H-20160121.aac* in that development dataset within the RTVE22 data was decoded without *−−accurate* option, due to its lower performance on the development data. Therefore, this development file was processed with the following default configuration setup: best_of=none, beam_search=none, temperature_increment_on_fallback=0.0.

6.1 ASR results

The performance of any STD system is highly influenced by the ASR system performance itself, especially for systems based on word recognition. Therefore, Table 6 presents the word error rate (WER) of the Whisper-based ASR system for all the datasets. It can be seen that the ASR system presents the highest error rates on the RTVE22 data. This is due to the *most difficult* speech present in those data (i.e., significant amount of overlapped speech, music, speech with music, etc.), which makes it more difficult to recognize. Overlapping speech is particularly harmful for ASR. This is a real RTVE22 data example of overlapped speech between speaker 1 and the initial turn of speaker 2 and the output produced by Whisper, where the overlapping is shown in bold font:

Speaker 1: Esa es la hamburguesa que se van a comer.

Speaker 2: **Los concursantes de hoy** esta es una de las hamburguesas que se van a comer.

Whisper output: Esa es la hamburguesa que se van a comer esta es una de las hamburguesas que se van a comer.

On the other hand, the ASR system exhibits the best performance on SPARL22 data, which present *more clean* speech from parliament sessions. It can also be seen that the results obtained on MAVIR data are worse than those on the SPARL22 data due to the more spontaneous speech present on MAVIR data but are consistently better than the results on RTVE22 data since RTVE22 data present the most difficult speech conditions. In any case, the performance of Whisper on these datasets is remarkably good, especially taking into account that neither adaptation nor fine-tuning have been applied.

Table 7 STD results for MAVIR development data

System	MTWV	ATWV	p(FA)	p(Miss)	Decision score
Whisper (single)	0.8115	0.8115	0.00002	0.169	0.055
Whisper (multi)	0.8814	0.8814	0.00002	0.099	0.055

MTWV maximum term weighted value, *ATWV* actual term weighed value, *FA* false alarm

Table 8 STD results for RTVE22 development data

System	MTWV	ATWV	p(FA)	p(Miss)	Decision score
Whisper (single)	0.7317	0.7317	0.00000	0.265	0.000
Whisper (multi)	0.7559	0.7559	0.00000	0.240	0.002

MTWV maximum term weighted value, *ATWV* actual term weighed value, *FA* false alarm

6.2 STD results

6.2.1 Development data

For development data, results are presented in Tables 7 and 8 for MAVIR and RTVE22 databases, respectively, for both the system that aims at single-word term

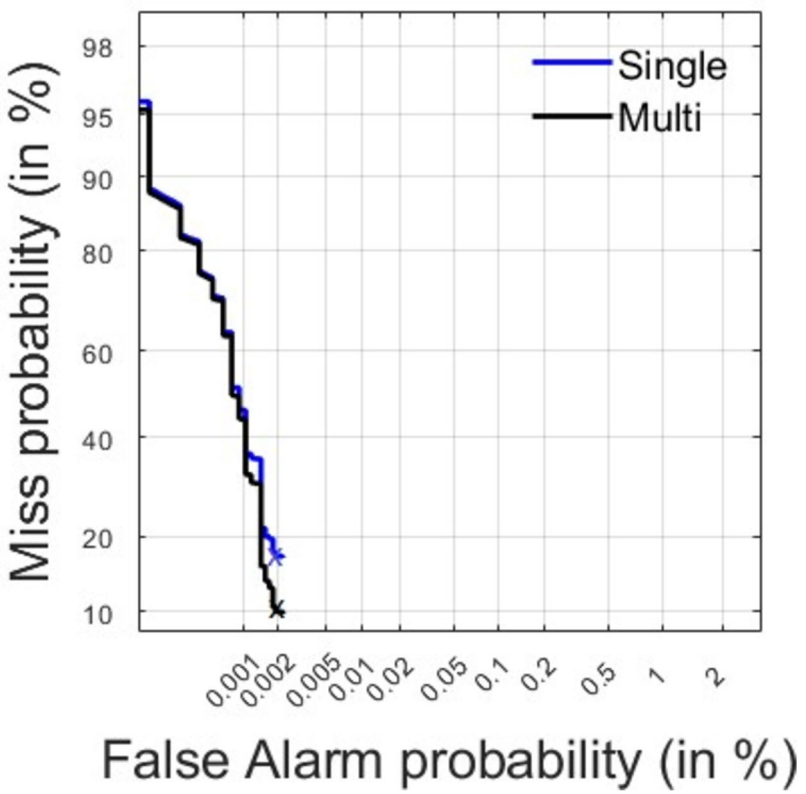

Fig. 5 DET curves for the MAVIR development data. Single refers to the *Whisper (single)* system and Multi refers to the *Whisper (multi)* system. X represents the operating point given the system decision threshold from which the ATWV is computed

detection (*Whisper (single)*) and the advanced system with the multi-word term detection capability (*Whisper (multi)*). The results show that incorporating the multi-word term detection on the Whisper-based approach does consistently improve the system performance. This improvement is statistically significant for a paired t-test ($p < 0.0001$). This is due to the fact that more terms are able to be detected with the enhanced capability of the *Whisper (multi)* system. From a numerical perspective, it can be seen that the *Whisper (multi)* system keeps the same number of FAs that of the *Whisper (single)* system and increases the number of hits for the MAVIR database, whereas for the RTVE22 database an increase in the number of hits and FAs is observed. This causes a higher improvement in the MTWV/ATWV figures on the MAVIR database. The higher gain of the *Whisper (multi)* system on MAVIR data with respect to that of RTVE22 data is due to the fact that RTVE22 data present more difficulties for ASR, which increases the FAs, especially for terms such as *partido popular* and *servicios de inteligencia*, which convey most of the multi-word term detection errors, since they are sometimes wrongly pronounced by the speaker. This is in fact has led to insertion errors in the ASR subsystem.

The DET curves for the development data are presented in Figs. 5 and 6 for MAVIR and RTVE22 data, respectively. For MAVIR data, the *Whisper (multi)* system does generally perform better than the *Whisper (single)* system, since more terms are able to be detected. For RTVE22 data, the *Whisper (single)* system performs the best for high and medium miss rates and the contrary occurs for low miss rates. This matches with the ATWV operating point, since this occurs at a low miss rate value. The errors mentioned before when incorporating the multi-word term detection capability make the *Whisper (single)* system curve is better than the *Whisper (multi)* system curve for most of the miss/FA ratios.

6.2.2 Test data

For test data, single-word and multi-word term detection results are presented in Tables 9, 10, and 11 for MAVIR, RTVE22 databases, and SPARL22 databases, respectively. These results do match with the development results. On the one hand, the *Whisper (multi)* system performs better

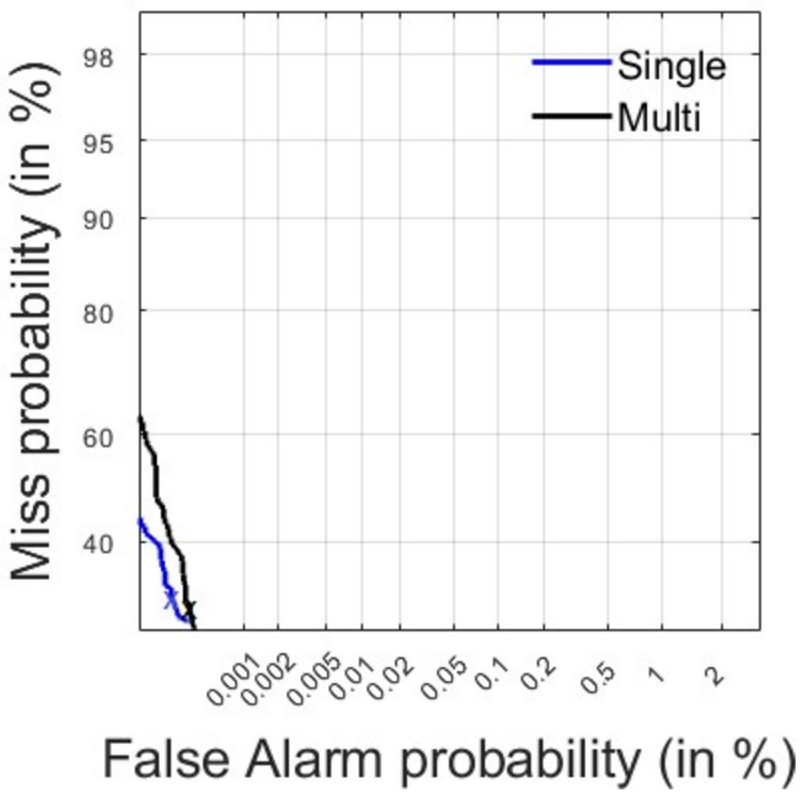

Fig. 6 DET curves for the RTVE22 development data. Single refers to the *Whisper (single)* system and Multi refers to the *Whisper (multi)* system. X represents the operating point given the system decision threshold from which the ATWV is computed

Table 9 STD results for MAVIR test data

System	MTWV	ATWV	p(FA)	p(Miss)	Decision score
Whisper (single)	0.7897	0.7886	0.00004	0.173	0.001
Whisper (multi)	0.8696	0.8685	0.00004	0.093	0.001

MTWV maximum term weighted value, *ATWV* actual term weighed value, *FA* false alarm

Table 10 STD results for RTVE22 test data

System	MTWV	ATWV	p(FA)	p(Miss)	Decision score
Whisper (single)	0.5558	0.5493	0.00001	0.433	0.006
Whisper (multi)	0.6414	0.6349	0.00001	0.347	0.006

MTWV maximum term weighted value, *ATWV* actual term weighed value, *FA* false alarm

Table 11 STD results for SPARL22 test data

System	MTWV	ATWV	p(FA)	p(Miss)	Decision score
Whisper (single)	0.7626	0.7626	0.00002	0.219	0.001
Whisper (multi)	0.8285	0.8285	0.00002	0.153	0.001

MTWV maximum term weighted value, *ATWV* actual term weighed value, *FA* false alarm

than the *Whisper (single)* system for all the databases, due to the enhanced term detection capability. These improvements are statistically significant for a paired t-test ($p < 0.02$). To explain that best performance of the *Whisper (multi)* system from a numerical perspective, it can be seen that the *Whisper (multi)* system maintains the same number of FAs as that of the *Whisper (single)* system and increases the number of hits for all the databases. This confirms the power of the multi-word term detector presented in this work.

The systems perform the worst on RTVE22 data, which present the most difficult speech. ATWV figures on MAVIR data are better than on SPARL22 data, while the contrary occurs in the ASR experiments. This may be due to the fact that the selected list of term produces less confusion in the ASR decoding (there are less insertion/deletion/substitution errors on the selected list of terms with respect to the other words in the speech data), so that the words that contribute to the higher WER are the rest (i.e., other common and stop words). On the other hand, the similar MTWV and ATWV for all the datasets does suggest that the detection threshold has been well calibrated by keeping all the occurrences provided by the ASR subsystem as actual STD detections. It is important

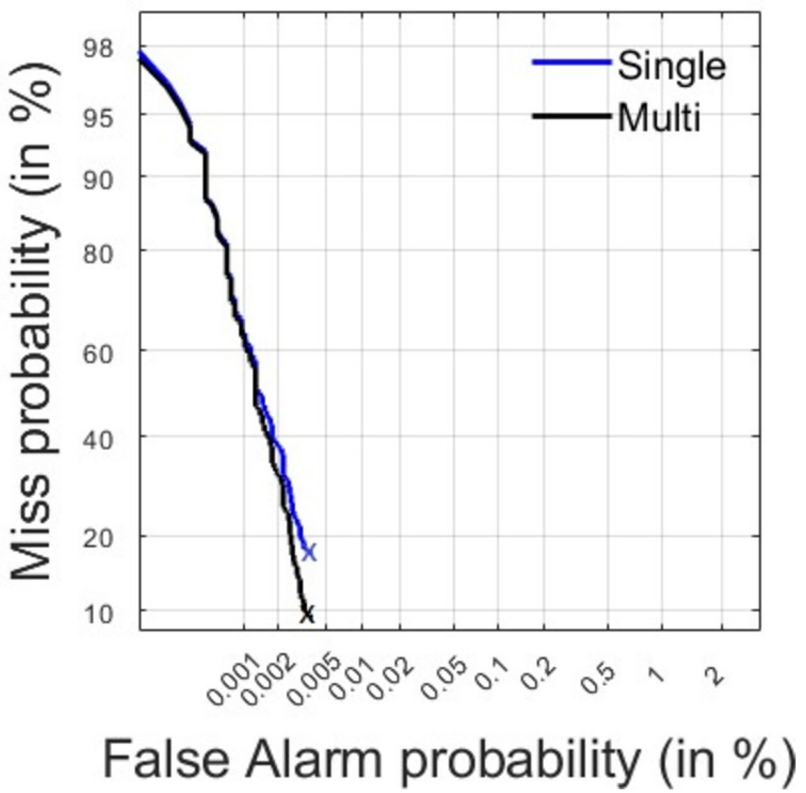

Fig. 7 DET curves for the MAVIR test data. Single refers to the *Whisper (single)* system and Multi refers to the *Whisper (multi)* system. X represents the operating point given the system decision threshold from which the ATWV is computed

to highlight the same MTWV and ATWV figures on the SPARL22 database. This is due to the combination of both low number of FAs and misses given by the STD systems, so that keeping all the detections as actual occurrences provides the same MTWV/ATWV figures.

The set of terms corresponding to the MAVIR database is the same as that of previous STD tasks of the SoS ALBAYZIN evaluations. Therefore, a fair comparison between all the systems submitted to the previous ALBAYZIN evaluations can be effectively carried out. The best system submitted so far to the STD task in SoS ALBAYZIN evaluations obtained an ATWV = 0.5724, which corresponds to a Kaldi-based approach that combined word lattices for in-vocabulary term detection and phoneme lattices for out-of-vocabulary term detection and was submitted to the STD task in 2016. The results obtained with the *Whisper (multi)* system are much better, since this obtains an ATWV = 0.8685. This improvement is statistically significant for a paired t-test ($p < 0.0001$). This shows the potential of the Whisper approach for the STD task.

In addition, the terms corresponding to the SPARL22 database include all the terms employed in the STD task of the previous SoS ALBAYZIN evaluation held in 2020 for that database. Therefore, when comparing the best result obtained in that evaluation on SPARL20 test data (ATWV = 0.5090) with the results obtained by the *Whisper (multi)* system on the common set of terms (ATWV = 0.8111), the power of the Whisper system when addressing term search is confirmed. This best performance is statistically significant for a paired t-test ($p < 0.0001$) as well.

The DET curves for the test data are presented in Figs. 7, 8, and 9 for MAVIR, RTVE22, and SPARL22 data, respectively. They also show that, in general, the *Whisper (multi)* system presents the best figures along different miss/FA ratios.

7 Additional analyses

Additional analyses have been carried out based on some term-related properties on the test data for the challenge databases. These term properties include in-language and out-of-language term detection, and single and multi-word term detection.

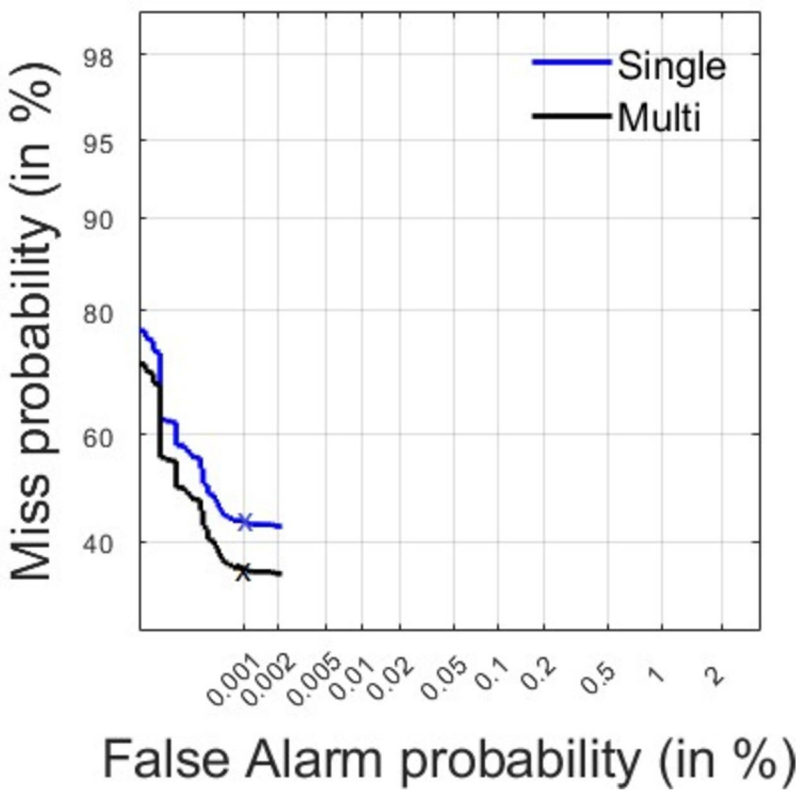

Fig. 8 DET curves for the RTVE22 test data. Single refers to the *Whisper (single)* system and Multi refers to the *Whisper (multi)* system. X represents the operating point given the system decision threshold from which the ATWV is computed

7.1 In-language vs. out-of-language term detection

Results for the test data by analyzing each system performance on Spanish (i.e., in-language) and foreign (i.e., out-of-language) terms are presented in Tables 12, 13, and 14 for MAVIR, RTVE22, and SPARL22 databases, respectively. These results show that INL terms are easier to detect, since the term target language matches with that of the model used when running the ASR system. However, the system performance degrades on OOL terms, since the language does not match with that of the model in the ASR system. The *Whisper (multi)* system outperforms the *Whisper (single)* system for all the datasets and term conditions, except for the SPARL22 OOL term detection, where the system performance remains the same. This is due to the fact that there are no OOL multi-word terms on these data.

7.2 Single-word vs. multi-word term detection

Results for the test data by analyzing each system performance on single-word and multi-word terms are presented in Tables 15, 16, and 17 for MAVIR, RTVE22, and SPARL22 databases, respectively. They show that multi-word terms are just able to be detected with the enhanced system (i.e., *Whisper (multi)*) and that the system performance on those terms is worse than on single-word terms, due to the inherent difficulty in detecting terms with more words (i.e., the ASR system must correctly recognize all the words that compose the multi-word term).

8 Conclusions and future work

Spoken term detection technology can be effectively applied when searching on speech content. STD evaluations held around the world provide a fair mechanism with which research groups and companies can effectively compare their system performance on a common framework. This paper presents two STD systems based on the Whisper ASR system aiming to find a set of terms in speech content and evaluates it on a Spanish STD evaluation challenge held as part of the ALBAYZIN evaluations from three different databases.

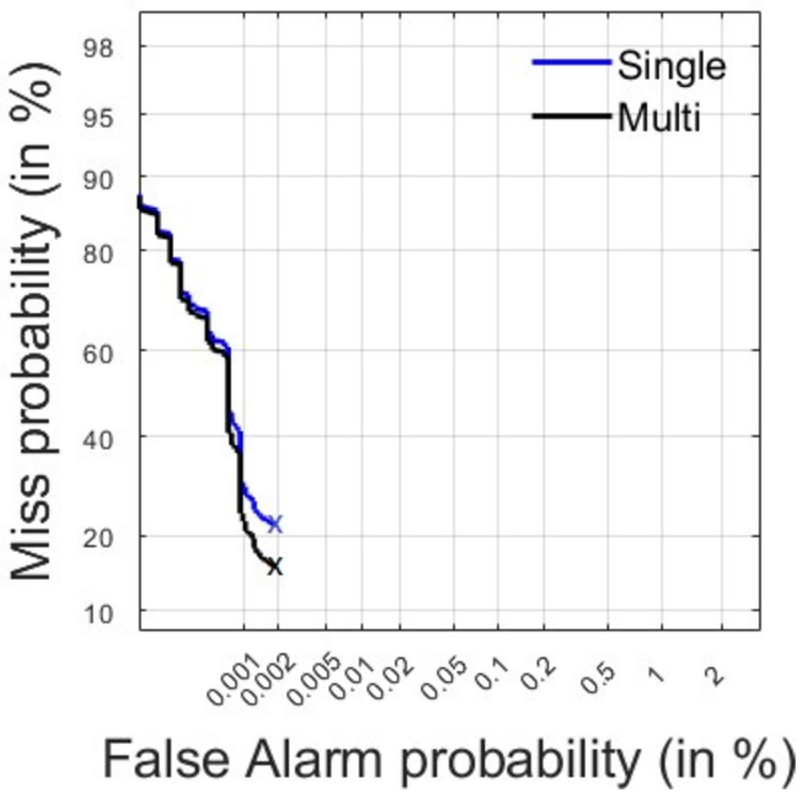

Fig. 9 DET curves for the SPARL22 test data. Single refers to the *Whisper (single)* system and Multi refers to the *Whisper (multi)* system. X represents the operating point given the system decision threshold from which the ATWV is computed

Table 12 STD results for INL and OOL term detection for MAVIR test data

	INL				OOL			
System	**MTWV**	**ATWV**	**p(FA)**	**p(Miss)**	**MTWV**	**ATWV**	**p(FA)**	**p(Miss)**
Whisper (single)	0.8170	0.8159	0.00003	0.150	0.4405	0.4177	0.00006	0.495
Whisper (multi)	0.8881	0.8870	0.00003	0.079	0.6405	0.6177	0.00006	0.295

INL in-language, *OOL* out-of-language, *MTWV* maximum term weighted value, *ATWV* actual term weighed value, *FA* false alarm

Table 13 STD results for INL and OOL term detection for RTVE22 test data

	INL				OOL			
System	**MTWV**	**ATWV**	**p(FA)**	**p(Miss)**	**MTWV**	**ATWV**	**p(FA)**	**p(Miss)**
Whisper (single)	0.6054	0.5947	0.00001	0.381	0.4326	0.4326	0.00000	0.566
Whisper (multi)	0.6882	0.6775	0.00001	0.298	0.5255	0.5255	0.00000	0.473

INL in-language, *OOL* out-of-language, *MTWV* maximum term weighted value, *ATWV* actual term weighed value, *FA* false alarm

Results show that the Whisper-based STD approach that integrates multi-word term detection capability does perform the best compared to all the systems submitted to all previous ALBAYZIN STD challenges held biannually since 2012.

Table 14 STD results for INL and OOL term detection for SPARL22 test data

	INL				OOL			
System	**MTWV**	**ATWV**	**p(FA)**	**p(Miss)**	**MTWV**	**ATWV**	**p(FA)**	**p(Miss)**
Whisper (single)	0.7763	0.7763	0.00002	0.203	0.5897	0.5897	0.00000	0.410
Whisper (multi)	0.8474	0.8474	0.00002	0.132	0.5897	0.5897	0.00000	0.410

INL in-language, *OOL* out-of-language, *MTWV* maximum term weighted value, *ATWV* actual term weighed value, *FA* false alarm

Table 15 STD results for single-word and multi-word terms for MAVIR test data

	Single				Multi			
System	**MTWV**	**ATWV**	**p(FA)**	**p(Miss)**	**MTWV**	**ATWV**	**p(FA)**	**p(Miss)**
Whisper (single)	0.8734	0.8722	0.00004	0.086	N/A	N/A	N/A	N/A
Whisper (multi)	0.8734	0.8722	0.00004	0.086	0.8333	0.8333	0.00000	0.167

Single single-word terms, *Multi* multi-word terms, *MTWV* maximum term weighted value, *ATWV* actual term weighed value, *FA* false alarm

Table 16 STD results for single-word and multi-word terms for RTVE22 test data

	Single				Multi			
System	**MTWV**	**ATWV**	**p(FA)**	**p(Miss)**	**MTWV**	**ATWV**	**p(FA)**	**p(Miss)**
Whisper (single)	0.6433	0.6358	0.00001	0.343	N/A	N/A	N/A	N/A
Whisper (multi)	0.6433	0.6358	0.00001	0.343	0.6294	0.6294	0.00000	0.371

Single single-word terms, *Multi* multi-word terms, *MTWV* maximum term weighted value, *ATWV* actual term weighed value, *FA* false alarm

Table 17 STD results for single-word and multi-word terms for SPARL22 test data

	Single				Multi			
System	**MTWV**	**ATWV**	**p(FA)**	**p(Miss)**	**MTWV**	**ATWV**	**p(FA)**	**p(Miss)**
Whisper (single)	0.8294	0.8294	0.00002	0.150	N/A	N/A	N/A	N/A
Whisper (multi)	0.8294	0.8294	0.00002	0.150	0.8182	0.8182	0.00000	0.182

Single single-word terms, *Multi* multi-word terms, *MTWV* maximum term weighted value, *ATWV* actual term weighed value, *FA* false alarm

The best system presented in this paper was developed aiming to be released as a baseline for the upcoming SoS 2024 ALBAYZIN evaluation challenge to encourage participants to improve the figures obtained in this paper. Improvements can be carried out from different methods: (1) constructing token lattices or increasing the beam search width in the ASR subsystem, (2) fine-tuning the ASR subsystem with a development set, by selecting the optimal parameter configuration in the Whisper_timestamped tool, and (3) employing more advanced threshold calibration approaches than that used in this work in the decision maker, among others. The baseline system can be found at https://github.com/javiertejedornoguerales/Whisper_STD.

Acknowledgements
This work has been partially funded by projects RTI2018-098091-B-I00 and PID2021-125943OB-I00 (Spanish Ministry of Science and Innovation and ERDF). Authors thank to Vivolab research group (University of Zaragoza) and Radio Televisión Española (RTVE) for providing the RTVE22 database. Authors also thank to Red Temática en Tecnologías del Habla (RTTH) for their support in ALBAYZIN evaluations and IberSPEECH 2022 organization committee for hosting the ALBAYZIN evaluations.

Authors' contributions
JT and DTT carried out the design of the search on speech challenge and labeling of the required data. JT built the ASR and STD systems. JT and DTT analyzed and discussed the results. All authors read and approved the final manuscript.

Funding
Funded by projects RTI2018-098091-B-I00 and PID2021-125943OB-I00 (Spanish Ministry of Science and Innovation and ERDF).

Availability of data and materials

The RTVE database is freely available subject to the terms of a license agreement with RTVE (http://catedrartve.unizar.es/rtvedatabase.html). Requirements for downloading the MAVIR database can be found in http://cartago.lllf.uam.es/mavir/index.pl?m=descargas. For details on SPARL22 database access, please contact Javier Tejedor (javier.tejedornoguerales@ceu.es).

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 17 July 2023 Accepted: 13 February 2024
Published online: 29 February 2024

References

1. K. Ng, V.W. Zue, Subword-based approaches for spoken document retrieval. Speech Comm. **32**(3), 157–186 (2000)
2. B. Chen, K.-Y. Chen, P.-N. Chen, Y.-W. Chen, Spoken document retrieval with unsupervised query modeling techniques. IEEE Trans. Audio Speech Lang. Process. **20**(9), 2602–2612 (2012)
3. T.-H. Lo, Y.-W. Chen, K.-Y. Chen, H.-M. Wang, B. Chen, in *Proceedings of ASRU*. Neural relevance-aware query modeling for spoken document retrieval. IEEE, Okinawa (2017), pp. 466–473
4. W.F.L. Heeren, F.M.G. Jong, L.B. Werff, M.A.H. Huijbregts, R.J.F. Ordelman, in *Proceedings of LREC*. Evaluation of spoken document retrieval for historic speech collections (2008), pp. 2037–2041
5. Y.-C. Pan, H.-Y. Lee, L.-S. Lee, Interactive spoken document retrieval with suggested key terms ranked by a Markov decision process. IEEE Trans. Audio Speech Lang. Process. **20**(2), 632–645 (2012)
6. Y.-W. Chen, K.-Y. Chen, H.-M. Wang, B. in *Proceedings of Interspeech*. Chen, Exploring the use of significant words language modeling for spoken document retrieval. ISCA, Stockholm (2017), pp. 2889–2893
7. A. Gupta, D. Yadav, A novel approach to perform context-based automatic spoken document retrieval of political speeches based on wavelet tree indexing. Multimed. Tools Appl. **80**, 22209–22229 (2021)
8. S.-W. Fan-Jiang, T.-H. Lo, B. Chen, in *Proceedings of ICASSP*. Spoken document retrieval leveraging BERT-based modeling and query reform ulation. IEEE, Barcelona (2020), pp. 8144–8148
9. H.-Y. Lin, T.-H. Lo, B. Chen, in *Proceedings ASRU*. Enhanced BERT-based ranking models for spoken document retrieval. IEEE, Sentosa (2019), pp. 601–606
10. Z.-Y. Wu, L.-P. Yen, K.-Y. Chen, in *Proceedings of ICASSP*. Generating pseudo-relevant representations for spoken document retrieval. ISCA, Brighton (2019), pp. 7370–7374
11. L.-P. Yen, Z.-Y. Wu, K.-Y. Chen, in *Proceedings of ICASSP*. A neural document language modeling framework for spoken document retriev al. IEEE, Barcelona (2020), pp. 8139–8143
12. Y. Moriya, G.J.F. Jones, in *Proceedings of SLT*. Improving noise robustness for spoken content retrieval using semi-supervised ASR and N-best transcripts for BERT-based ranking models. IEEE, Doha (2023), pp. 398–405
13. E. Villatoro-Tello, S. Madikeri, P. Motlicek, A. Ganapathiraju, A.V. Ivanov, in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Expanded lattice embeddings for spoken document retrieval on informal meetings. ACM, Madrid (2022), pp. 2669–2674
14. P. Gao, J. Liang, P. Ding, B. Xu, in *Proceedings of ICASSP*. A novel phone-state matrix based vocabulary-independent keyword spotting method for spontaneous speech. IEEE, Honolulu (2007), pp. 425–428
15. A. Mandal, J. Hout, Y.-C. Tam, V. Mitra, Y. Lei, J. Zheng, D. Vergyri, L. Ferrer, M. Graciarena, A. Kathol, H. Franco, in *Proceedings of Interspeech*. Strategies for high accuracy keyword detection in noisy channels. ISCA, Lyon (2013), pp. 15–19
16. S. Panchapagesan, M. Sun, A. Khare, S. Matsoukas, A. Mandal, B. Hoffmeister, S. Vitaladevuni, in *Proceedings of Interspeech*. Multi-task learning and weighted cross-entropy for DNN-based keyword spotting. ISCA, San Francisco (2016), pp. 760–764
17. H. Mazzawi, X. Gonzalvo, A. Kracun, P. Sridhar, N. Subrahmanya, I.L. Moreno, H.J. Park, P. Violette, in *Proceedings of Interspeech*. Improving keyword spotting and language identification via Neural Architecture Search at Scale. ISCA, Graz (2019), pp. 1278–1282
18. T. Mo, Y. Yu, M. Salameh, D. Niu, S. Jui, in *Proceedings of Interspeech*. Neural architecture search for keyword spotting. ISCA, Shanghai (2020), pp. 1982–1986
19. H.-J. Park, P. Zhu, I.L. Moreno, N. Subrahmanya, in *Proceedings of Interspeech*. Noisy student-teacher training for robust keyword spotting. ISCA, Brno (2021), pp. 331–335
20. B. Wei, M. Yang, T. Zhang, X. Tang, X. Huang, K. Kim, J. Lee, K. Cho, S.-U. Park, in *Proceedings of Interspeech*, End-to-end transformer-based open-vocabulary keyword spotting with location-guided local attention. ISCA, Brno (2021), pp. 361–365
21. R. Kirandevraj, V.K. Kurmi, V. Namboodiri, C.V. Jawahar, in *Proceedings of Interspeech*. Generalized keyword spotting using ASR embeddings. ISCA , Incheon (2022), pp. 126–130
22. Z. Yang, S. Sun, J. Li, X. Zhang, X. Wang, L. Ma, L. Xie, in *Proceedings of Interspeech*. CaTT-KWS: A multi-stage customized keyword spotting framework based on cascaded transducer-transformer. ISCA, Incheon (2022), pp. 1681–1685
23. L. Lei, G. Yuan, H. Yu, D. Kong, Y. He, Multilingual customized keyword spotting using similar-pair contrastive learning. IEEE/ACM Trans. Audio Speech Lang. Process. **31**, 2437–2447 (2023)
24. M. Dampfhoffer, T. Mesquida, E. Hardy, A. Valentian, L. Anghel, in *Proceedings of ICASSP*, Leveraging sparsity with spiking recurrent neural networks for energy-efficient keyword spotting. IEEE, Rhodes island (2023), pp. 1–5
25. E. van der Westhuizen, H. Kamper, R. Menon, J. Quinn, T. Niesler, Feature learning for efficient ASR-free keyword spotting in low-resource languages. Comp. Speech Lang. **71**, 101275 (2022)
26. K. Ding, M. Zong, J. Li, B. Li, in *Proceedings of ICASSP*. Letr: A lightweight and efficient transformer for keyword spotting. IEEE, Singapore (2022), pp. 7987–7991
27. Z. Wang, L. Wan, B. Zhang, Y. Huang, S.-W. Li, M. Sun, X. Lei, Z. Yang, in *Proceedings of ICASSP*, Disentangled training with adversarial examples for robust small-footprint keyword spotting. IEEE, Rhodes island (2023), pp. 1–5
28. A. Buzo, H. Cucu, C. Burileanu, in *Proceedings of MediaEval*. SpeeD@ MediaEval 2014: Spoken term detection with robust multilingual phone recognition. MediaEval Multimedia, Barcelona (2014), pp. 721–722
29. R. Konno, K. Ouchi, M. Obara, Y. Shimizu, T. Chiba, T. Hirota, Y. Itoh, in *Proceedings of NTCIR-12*. An STD system using multiple STD results and multiple rescoring method for NTCIR-12 SpokenQuery &Doc task. National Institute of Informatics, Tokyo (2016), pp. 200–204
30. R. Jarina, M. Kuba, R. Gubka, M. Chmulik, M. Paralic, in *Proceedings of MediaEval*. UNIZA system for the spoken web search task at MediaEval 2013. MediaEval Multimedia, Barcelona (2013), pp. 791–792
31. X. Anguera, M. Ferrarons, in *Proceedings of ICME*. Memory efficient subsequence DTW for query-by-example spoken term detection. IEEE, San Jose (2013), pp. 1–6
32. C. Chan, L. Lee, in *Proceedings of Interspeech*. Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping. ISCA, Chiba (2010), pp. 693–696
33. J. Huang, W. Gharbieh, Q. Wan, H.S. Shim, H.C. Lee, in *Proceedings of Interspeech*. QbyE-MLPMixer: Query-by-example open-vocabulary keyword spotting using MLPMixer. ISCA, Incheon (2022), pp. 5200–5204
34. S.-Y. Chang, G. Prakash, Z. Wu, T. Sainath, B. Li, Q. Liang, A. Stambler, S. Upadhyay, M. Faruqui, T. Strohman, in *Proceedings of Interspeech*. Streaming intended query detection using E2E modeling for continued conversation. ISCA, Incheon (2022), pp. 1826–1830
35. D. Ram, L. Miculicich, H. Bourlard, Neural network based end-to-end query by example spoken term detection. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 1416–1427 (2020)
36. J. Huang, W. Gharbieh, H.S. Shim, E. Kim, in *Proceedings of ICASSP*. Query-by-example keyword spotting system using multi-head attention and soft-triple loss. IEEE, Toronto (2021), pp. 6858–6862

37. D. Ram, L. Miculicich, H. Bourlard, in *Proceedings of ASRU*. Multilingual bottleneck features for query by example spoken term detection. IEEE, Sentosa (2019), pp. 621–628
38. Y. Hu, S. Settle, K. Livescu, in *Proceedings of SLT*. Acoustic span embeddings for multilingual query-by-example search. IEEE, Shenzhen (2021), pp. 935–942
39. Y. Yuan, L. Xie, C.-C. Leung, H. Chen, B. Ma, Fast query-by-example speech search using attention-based deep binary embeddings. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 1988–2000 (2020)
40. P.M. Reuter, C. Rollwage, B.T. Meyer, in *Proceedings of ICASSP*. Multilingual query-by-example keyword spotting with metric learning and phoneme-to-embedding mapping. IEEE, Rhodes island (2023), pp. 1–5
41. R. Khwildi, A.O. Zaid, F. Dufaux, Query-by-example HDR image retrieval based on CNN. Multimed. Tools Appl. **80**, 15413–15428 (2021)
42. P. Lopez-Otero, J. Parapar, A. Barreiro, Statistical language models for query-by-example spoken document retrieval. Multimedia Tools Appl. **79**, 7927–7949 (2020)
43. J. Mamou, B. Ramabhadran, O. Siohan, in *Proceedings of ACM SIGIR*. Vocabulary independent spoken term detection. ACM, Amsterdam (2007), pp. 615–622
44. J. Mamou, B. Ramabhadran, in *Proceedings of Interspeech*. Phonetic query expansion for spoken document retrieval. ISCA, Brisbane (2008), pp. 2106–2109
45. D. Can, E. Cooper, A. Sethy, C. White, B. Ramabhadran, M. Saraclar, in *Proceedings of ICASSP*. Effect of pronunciations on OOV queries in spoken term detection. IEEE, Taipei (2009), pp. 3957–3960
46. A. Rosenberg, K. Audhkhasi, A. Sethy, B. Ramabhadran, M. Picheny, in *Proceedings of ICASSP*. End-to-end speech recognition and keyword search on low-resource languages. IEEE, New Orleans (2017), pp. 5280–5284
47. K. Audhkhasi, A. Rosenberg, A. Sethy, B. Ramabhadran, B. Kingsbury, in *Proceedings of ICASSP*. End-to-end ASR-free keyword search from speech. IEEE, New Orleans (2017), pp. 4840–4844
48. K. Audhkhasi, A. Rosenberg, A. Sethy, B. Ramabhadran, B. Kingsbury, End-to-end ASR-free keyword search from speech. IEEE J. Sel. Top. Signal Process. **11**(8), 1351–1359 (2017)
49. J.G. Fiscus, J. Ajot, J.S. Garofolo, G. Doddington, in *Proceedings of SSCS*. Results of the 2006 spoken term detection evaluation. ACM, Amsterdam (2007), pp. 45–50
50. W. Hartmann, L. Zhang, K. Barnes, R. Hsiao, S. Tsakalidis, R. Schwartz, in *Proceedings of Interspeech*. Comparison of multiple system combination techniques for keyword spotting. ISCA, San Francisco (2016), pp. 1913–1917
51. T. Alumae, D. Karakos, W. Hartmann, R. Hsiao, L. Zhang, L. Nguyen, S. Tsakalidis, R. Schwartz, in *Proceedings of ICASSP*. The 2016 BBN Georgian telephone speech keyword spotting system. IEEE, New Orleans (2017), pp. 5755–5759
52. D. Vergyri, A. Stolcke, R.R. Gadde, W. Wang, in *Proceedings of NIST Spoken Term Detection Workshop (STD 2006)*. The SRI 2006 spoken term detection system. National Institute of Standards and Technolo gy, Gaithersburg (2006), pp. 1–15
53. D. Vergyri, I. Shafran, A. Stolcke, R.R. Gadde, M. Akbacak, B. Roark, W. Wang, in *Proceedings of Interspeech*. The SRI/OGI 2006 spoken term detection system. ISCA, Antwerp (2007), pp. 2393–2396
54. M. Akbacak, D. Vergyri, A. Stolcke, in *Proceedings of ICASSP*. Open-vocabulary spoken term detection using graphone-based hybrid recognition systems. IEEE, Las Vegas (2008), pp. 5240–5243
55. I. Szöke, M. Fapšo, M. Karafiát, L. F. Burget, 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. Springer, Brno (2008), pp. 237–247
56. I. Szöke, L. Burget, J. Černocký, M. Fapšo, in *Proceedings of SLT*. Sub-word modeling of out of vocabulary words in spoken term detection. IEEE, Goa, India (2008), pp. 273–276
57. I. Szöke, M. Fapšo, L. Burget, J. Černocký, in *Proceedings of Speech Search Workshop at SIGIR*. Hybrid word-subword decoding for spoken term detection. ACM, Singapore (2008), pp. 42–48
58. S. Meng, P. Yu, J. Liu, F. Seide, in *Proceedings of ICASSP*. Fusing multiple systems into a compact lattice index for Chinese spoken term detection. IEEE, Las Vegas (2008), pp. 4345–4348
59. S. Shah, S. Sitaram, in *Proceedings of International Conference on Data Mining*. Using monolingual speech recognition for spoken term detection in code-switched Hindi-English speech. IEEE, Beijing (2019), pp. 1–5
60. 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)
61. R. Wallace, R. Vogt, B. Baker, S. Sridharan, in *Proceedings of ICASSP*. Optimising figure of merit for phonetic spoken term detection. IEEE, Dallas (2010), pp. 5298–5301
62. C. Parada, A. Sethy, M. Dredze, F. Jelinek, in *Proceedings of Interspeech*. A spoken term detection framework for recovering out-of-vocabulary words using the web. ISCA, Chiba (2010), pp. 1269–1272
63. A. Jansen, K. Church, H. Hermansky, in *Proceedings of Interspeech*. Towards spoken term discovery at scale with zero resources. ISCA, Chiba (2010), pp. 1676–1679
64. C. Parada, A. Sethy, B. Ramabhadran, in *Proceedings of ICASSP*. Balancing false alarms and hits in spoken term detection. IEEE, Dallas (2010), pp. 5286–5289
65. J. Trmal, M. Wiesner, V. Peddinti, X. Zhang, P. Ghahremani, Y. Wang, V. Manohar, H. Xu, D. Povey, S. Khudanpur, in *Proceedings of Interspeech*. The Kaldi OpenKWS system: Improving low resource keyword search. ISCA, Stockholm (2017), pp. 3597–3601
66. D. Schneider, T. Mertens, M. Larson, J. Kohler, in *Proceedings of Interspeech*. Contextual verification for open vocabulary spoken term detection. ISCA, Chiba (2010), pp. 697–700
67. C.-A. Chan, L.-S. Lee, in *Proceedings of Interspeech*. Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping. ISCA, Chiba (2010), pp. 693–696
68. C.-P. Chen, H.-Y. Lee, C.-F. Yeh, L.-S. Lee, in *Proceedings of Interspeech*. Improved spoken term detection by feature space pseudo-relevance feedback. ISCA, Chiba (2010), pp. 1672–1675
69. P. Motlicek, F. Valente, P. Garner, in *Proceedings of Interspeech*. English spoken term detection in multilingual recordings. ISCA, Chiba (2010), pp. 206–209
70. J. Wintrode, J. Wilkes, in *Proceedings of ICASSP*. Fast lattice-free keyword filtering for accelerated spoken term detection. IEEE, Barcelona (2020), pp. 7469–7473
71. T.S. Fuchs, Y. Segal, J. Keshet, in *Proceedings of ICASSP*. CNN-based spoken term detection and localization without dynamic programming . IEEE, Toronto (2021), pp. 6853–6857
72. B. Yusuf, M. Saraclar, in *Proceedings of Interspeech*. An empirical evaluation of DTW subsampling methods for keyword search (2019), pp. 2673–2677
73. V.L.V. Nadimpalli, S. Kesiraju, R. Banka, R. Kethireddy, S.V. Gangashetty, Resources and benchmarks for keyword search in spoken audio from low-resource indian languages. IEEE Access **10**, 34789–34799 (2022)
74. NIST, The Spoken Term Detection (STD) 2006 Evaluation Plan (2006). https://catalog.ldc.upenn.edu/docs/LDC2011S02/std06-evalplan-v10.pdf. Accessed 26 Feb 2024
75. NIST, OpenKWS13 Keyword Search Evaluation Plan (National Institute of Standards and Technology (NIST), Gaithersburg, 2013). https://www.nist.gov/system/files/documents/itl/iad/mig/OpenKWS13-EvalPlan.pdf. Accessed 26 Feb 2024
76. NIST, Draft KWS14 Keyword Search Evaluation Plan (National Institute of Standards and Technology (NIST), Gaithersburg, 2013). https://www.nist.gov/system/files/documents/itl/iad/mig/KWS14-evalplan-v11.pdf. Accessed 26 Feb 2024
77. NIST, KWS15 Keyword Search Evaluation Plan (National Institute of Standards and Technology (NIST), Gaithersburg, 2015). https://www.nist.gov/system/files/documents/itl/iad/mig/KWS15-evalplan-v05.pdf. Accessed 26 Feb 2024
78. NIST, Draft KWS16 Keyword Search Evaluation Plan (National Institute of Standards and Technology (NIST), Gaithersburg, 2016). https://www.nist.gov/system/files/documents/itl/iad/mig/KWS16-evalplan-v04.pdf. Accessed 26 Feb 2024

79. Z. Lv, M. Cai, W.-Q. Zhang, J. Liu, in *Proceedings of Interspeech*. A novel discriminative score calibration method for keyword search. ISCA, San Francisco (2016), pp. 745–749
80. N.F. Chen, V.T. Pharri, H. Xu, X. Xiao, V.H. Do, C. Ni, I.-F. Chen, S. Sivadas, C.-H. Lee, E.S. Chng, B. Ma, H. Li, in *Proceedings of ICASSP*. Exemplar-inspired strategies for low-resource spoken keyword search in Swahili. IEEE, Shanghai (2016), pp. 6040–6044
81. C. Ni, C.-C. Leung, L. Wang, H. Liu, F. Rao, L. Lu, N.F. Chen, B. Ma, H. Li, in *Proceedings of ICASSP*. Cross-lingual deep neural network based submodular unbiased data selection for low-resource keyword search. IEEE, Shanghai (2016), pp. 6015–6019
82. M. Cai, Z. Lv, C. Lu, J. Kang, L. Hui, Z. Zhang, J. Liu, in *Proceedings of ASRU*. High-performance swahili keyword search with very limited language pack: The THUEE system for the OpenKWS15 evaluation. IEEE, Scottsdale (2015), pp. 215–222
83. N.F. Chen, C. Ni, I.-F. Chen, S. Sivadas, V.T. Pham, H. Xu, X. Xiao, T.S. Lau, S.J. Leow, B.P. Lim, C.-C. Leung, L. Wang, C.-H. Lee, A. Goh, E.S. Chng, B. Ma, H. Li, in *Proceedings of ICASSP*. Low-resource keyword search strategies for Tamil. IEEE, South Brisbane (2015), pp. 5366–5370
84. L. Mangu, G. Saon, M. Picheny, B. Kingsbury, in *Proceedings of ICASSP*, Order-free spoken term detection. IEEE, South Brisbane (2015), pp. 5331–5335
85. C. Heerden, D. Karakos, K. Narasimhan, M. Davel, R. Schwartz, in *Proceedings of ICASSP*. Constructing sub-word units for spoken term detection. IEEE, South Brisbane (2017), pp. 5780–5784
86. W. Hartmann, D. Karakos, R. Hsiao, L. Zhang, T. Alumae, S. Tsakalidis, R. Schwartz, in *Proceedings of ICASSP*. Analysis of keyword spotting performance across IARPA babel languages. ISCA, New Orleans (2017), pp. 5765–5769
87. C. Ni, C.-C. Leung, L. Wang, N.F. Chen, B. Ma, in *Proceedings of ICASSP*. Efficient methods to train multilingual bottleneck feature extractors for low resource keyword search. ISCA, New Orleans (2017), pp. 5650–5654
88. A. Ragni, D. Saunders, P. Zahemszky, J. Vasilakes, M.J.F. Gales, K.M. Knill, in*Proceedings of ICASSP*. Morph-to-word transduction for accurate and efficient automatic speech recognition and keyword search. ISCA, New Orleans (2017), pp. 5770–5774
89. X. Chen, A. Ragnil, J. Vasilakes, X. Liu, K. Knilll, M.J..F. Gales, in *Proceedings of ICASSP*. Recurrent neural network language models for keyword search. ISCA, New Orleans (2017), pp. 5775–5779
90. V.T. Pham, H. Xu, X. Xiao, N.F. Chen, E.S. Chng, in *Proceedings of International Symposium on Information and Communication Technology*. Pruning strategies for partial search in spoken term detection. ACM, Nha Trang (2017), pp. 114–119
91. V.T. Pham, H. Xu, X. Xiao, N.F. Chen, E.S. Chng, Re-ranking spoken term detection with acoustic exemplars of keywords. Speech Comm. **104**, 12–23 (2018)
92. R. Lileikyte, T. Fraga-Silva, L. Lamel, J.-L. Gauvain, A. Laurent, G. Huang, in *Proceedings of ICASSP*. Effective keyword search for low-resourced conversational speech. ISCA, New Orleans (2017), pp. 5785–5789
93. Y. Khokhlov, I. Medennikov, A. Romanenko, V. Mendelev, M. Korenevsky, A. Prudnikov, N. Tomashenko, A. Zatvornitsky, in *Proceedings of Interspeech*. The STC keyword search system for OpenKWS 2016 evaluati on. ISCA, Stockholm (2017), pp. 3602–3606
94. T. Sakai, H. Joho, in *Proceedings of NTCIR-9*. Overview of NTCIR-9. National Institute of Informatics, Tokyo (2011), pp. 1–7
95. T. Akiba, H. Nishizaki, K. Aikawa, X. Hu, Y. Itoh, T. Kawahara, S. Nakagawa, H. Nanjo, Y. Yamashita, in *Proceedings of NTCIR-10*. Overview of the NTCIR-10 SpokenQueryDoc-2 task. National Institute of Informatics, Tokyo (2013), pp. 1–15
96. T. Akiba, H. Nishizaki, H. Nanjo, G.J.F. Jones, in *Proceedings of NTCIR-11*. Overview of the NTCIR-11 SpokenQuery &Doc task. National Institute of Informatics, Tokyo (2014), pp. 1–15
97. T. Akiba, H. Nishizaki, H. Nanjo, G.J.F. Jones, in *Proceedings of NTCIR-12*. Overview of the NTCIR-12 SpokenQuery &Doc-2 task. National Institute of Informatics, Tokyo (2016), pp. 1–13
98. J. Wang, Y. He, C. Zhao, Q. Shao, W.-W. Tu, T. Ko, H.-y. Lee, L. Xie, in *Proceedings of Interspeech*. Auto-KWS 2021 challenge: Task, datasets, and baselines. ISCA, Brno (2021), pp. 4244–4248
99. J. Tejedor, D.T. Toledano, P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, A. Cardenal, J.D. Echeverry-Correa, A. Coucheiro-Limeres, J. Olcoz, A. Miguel, Spoken term detection ALBAYZIN 2014 evaluation: Overview, systems, results, and discussion. EURASIP J. Audio Speech Music Process. **2015**(21), 1–27 (2015)
100. J. Tejedor, D.T. Toledano, P. Lopez-Otero, L. Docio-Fernandez, L. Serrano, I. Hernaez, A. Coucheiro-Limeres, J. Ferreiros, J. Olcoz, J. Llombart, ALBAYZIN 2016 spoken term detection evaluation: An international open competitive evaluation in Spanish. EURASIP J. Audio Speech Music Process. **2017**(22), 1–23 (2017)
101. J. Tejedor, D.T. Toledano, P. Lopez-Otero, L. Docio-Fernandez, A.R. Montalvo, J.M. Ramirez, M. Peñagarikano, L.-J. Rodriguez-Fuentes, ALBAYZIN 2018 spoken term detection evaluation: A multi-domain international evaluation in Spanish. EURASIP J. Audio Speech Music Process. **2019**(16), 1–37 (2019)
102. J. Tejedor, D.T. Toledano, J.M. Ramirez, A.R. Montalvo, J.I. Alvarez-Trejos, The multi-domain international search on speech 2020 ALBAYZIN evaluation: Overview, systems, results, discussion and post-evaluation analyses. Appl. Sci. **11**(18), 8519 (2021)
103. A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision (2022). arXiv preprint arXiv:2212.04356
104. A.M. Sandoval, L.C. Llanos, in *Proceedings of Iberspeech*. MAVIR: A corpus of spontaneous formal speech in Spanish and English. RTTH, Madrid (2012)
105. E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Perez, A. Prada, RTVE 2018, 2020 and 2022 Database Description (Vivolab, Aragon Institute for Engineering Resarch (I3A), University of Zaragoza, Spain, 2022). https://catedrartve.unizar.es/reto2022/RTVE2022DB.pdf. Accessed 26 Feb 2024
106. A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, in *Proceedings of Eurospeech*. The DET curve in assessment of detection task performance. ISCA, Rhodes (1997), pp. 1895–1898
107. NIST, Evaluation Toolkit (STDEval) Software (National Institute of Standards and Technology (NIST), Gaithersburg, 1996). https://www.nist.gov/itl/iad/mig/tools. Accessed 26 Feb 2024
108. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need. Adv. Neural Inf. Process. Syst. **30**, 1–11 (2017)
109. P. Gage, A new algorithm for data compression. C Users J. **12**(2), 23–38 (1994)
110. R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units (2015). arXiv preprint arXiv:1508.07909
111. J. Louradour, Whisper-timestamped (GitHub, 2023)
112. T. Giorgino, Computing and visualizing dynamic time warping alignments in r: The dtw package. J. Stat. Softw. **31**(7), 1–24 (2009)
113. J.G. Fiscus, J. Ajot, J.S. Garofolo, G. Doddington, in *Proceedings of SIGIR Workshop Searching Spontaneous Conversational Speech*. Results of the 2006 spoken term detection evaluation. ACM, Amsterdam (2007), pp. 45–50

Publisher's Note