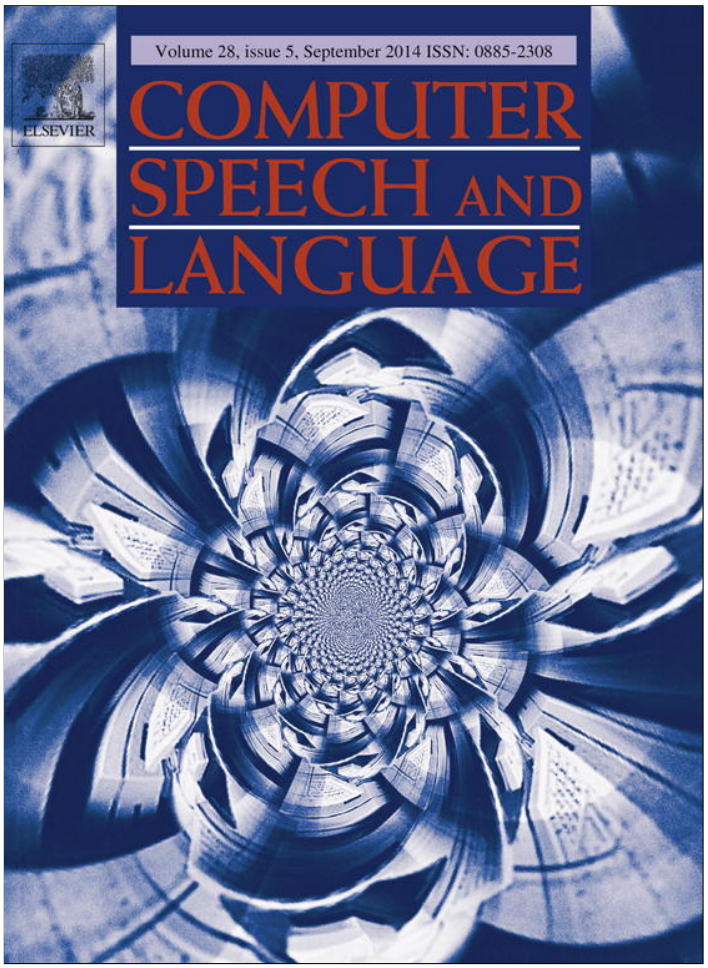

Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 28 (2014) 1066–1082

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Language independent search in MediaEval's Spoken Web Search task

Florian Metze [a,*], Xavier Anguera [c], Etienne Barnard [b], Marelie Davel [b], Guillaume Gravier [d]

[a] *Carnegie Mellon University, Pittsburgh, PA, USA*
[b] *North-West University, Vanderbijlpark, South Africa*
[c] *Telefonica Research, Barcelona, Spain*
[d] *CNRS–IRISA, Rennes, France*

Received 2 August 2012; received in revised form 13 October 2013; accepted 29 December 2013
Available online 27 January 2014

Abstract

In this paper, we describe several approaches to language-independent spoken term detection and compare their performance on a common task, namely "Spoken Web Search". The goal of this part of the MediaEval initiative is to perform low-resource language-independent audio search using audio as input. The data was taken from "spoken web" material collected over mobile phone connections by IBM India as well as from the LWAZI corpus of African languages. As part of the 2011 and 2012 MediaEval benchmark campaigns, a number of diverse systems were implemented by independent teams, and submitted to the "Spoken Web Search" task. This paper presents the 2011 and 2012 results, and compares the relative merits and weaknesses of approaches developed by participants, providing analysis and directions for future research, in order to improve voice access to spoken information in low resource settings.

Keywords: Low-resource speech technology; Evaluation; Spoken web; Spoken term detection

1. Introduction

In recent years, speech technology has emerged as an enabling technology for increasing the accessibility of information for a number of quite diverse use cases. These include searching large archives of audio-visual material, dialog systems for access to personal information and (mobile) web search, as well as applications in language learning and pronunciation training. A particularly deserving aspect of these is the potential of speech technologies to foster participation of disabled, low-literate, or "minority" users in the information society.

The last case has proven to be particularly challenging, because resources of any kind are usually scarce for minority languages, dialects and other non-mainstream conditions, which can therefore not be approached with the typical "there is no data like more data" engineering approach. Clearly, society would benefit greatly from the ability to easily process

* Corresponding author. Tel.: +1 412 2688984.
E-mail address: fmetze@cs.cmu.edu (F. Metze).

0885-2308/$ – see front matter
http://dx.doi.org/10.1016/j.csl.2013.12.004

audio in any language (or dialect), or language independently, without having to spend resources on language-specific development, but significant research is still needed in that area.

"Spoken Web Search" involves searching *for* audio content, *within* audio content, *using* an audio query, in a language, dialect, or domain for which only very limited resources are available. The original motivation for this task was to be able to provide voice-based access to spoken documents created by local community efforts in rural India. Because no experts are available to create and maintain dedicated speech dialog systems, acoustic similarity and keyword search could be used to access information. A caller would for example say *"weather tomorrow"* and retrieve a spoken document which contains the phrase *"the weather tomorrow will be . . ."* (in a matching dialect). While such a retrieval-based approach is clearly limited when compared to fully developed dialog systems, it is still preferable to not having any capability to access information at all. The fact that users in such applications will often be repeat callers, and therefore will be familiar with the system, also enhances the potential efficacy of this approach.

The main challenge is therefore to develop approaches to spoken term detection (rather than full speech-to-text) that scale to many languages, dialects, and domains very quickly, without requiring data and language or technology experts. An efficient large-scale deployment with many users is desirable, but not the primary goal. To solve this problem, two research avenues present themselves: port existing speech recognition approaches, or build dedicated solutions. When starting from existing speech recognition approaches and resources, techniques have to be found which will make them useful in low-resource settings. Multi-lingual modeling and cross-lingual (or -dialectal) transfer have been used in the past to do that. As an alternative, limited keyword search or acoustic pattern matching systems can be tailored specifically to the target use case. It may therefore be possible to develop them with relatively little data, or even zero (outside) resources.

To compare these two approaches, and analyze the trade-offs entailed by such a design decision, "Spoken Web Search" (SWS) was run as a challenge-style task at MediaEval 2011 (Rajput and Metze, 2011) and MediaEval 2012 (Metze et al., 2012). This evaluation attempts to provide a common evaluation corpus and baseline for research on language-independent search and retrieval of real-world speech data, with a special focus on low-resource languages.

In Section 2, we survey the field of low-resource acoustic pattern matching and spoken term detection. Section 3 presents the "Spoken Web Search" task as a public data set and evaluation campaign to initiate research and discussion in this research area. A unified view and discussion of the approaches implemented by the participants is given in Sections 4–7, following the different steps of a typical system. In Section 8, we discuss results achieved in the 2011 and 2012 evaluations, and provide research directions for the future.

2. Related work

The World Wide Web has changed the information landscape for the developed world, where citizens are used to accessing information about almost anything, anytime and on any device. The Web 2.0 has democratized the web further by enabling user-generated content through wikis, blogs, and more recently through social networking websites. However, the developing regions still face several challenges in being part of this information revolution: low literacy (Education, 2010) and lack of internet penetration (Internet, 2010) being some of them.

On the other hand, India, China, Brazil and Indonesia taken together have more mobile phones than the Top 50 developed countries taken together (Internet, 2010). Voice-based information systems are therefore evolving as an alternative to the text/ visual web, and could potentially achieve significant penetration. Users in developing regions are now able to create, access and share content using just their voice and a phone. Several audio-based information systems have been deployed in the last five years. The Healthline (Sherwani et al., 2007) system provides reliable health information for community health workers in developing countries. The AudioWiki system (Kotkar et al., 2008) provides a repository of spoken content that can be modified through a low-cost phone and can support any language. Adoption issues have recently been studied in Raza et al. (2013). The Spoken Web system has been used in agriculture (Patel et al., 2010), employment (Kumar et al., 2008), social (Agarwal et al., 2009, 2010) and several other settings (Agarwal et al., 2010; Diao et al., 2010).

While these examples clearly illustrate the usefulness of a voice-based information system, they also pose challenging research problems. In some of these systems, most of the content is user-generated and in local languages and dialects. The content is therefore mostly spontaneous and colloquial in style, with a lot and a high variety of background noise. Voice being a sequential modality, navigation by commands is often a challenge, suggesting that audio content search be applied in these information systems. Leaving aside questions of usability and deployment, efficient keyword-style

search is a key requirement in order to create a usable and low-cost (for both the provider and consumer) audio information system for the target user group.

The "Spoken Web Search" task therefore sits at the intersection of two research domains that have recently seen significant activity, namely *speech technology for under-resourced languages*, as discussed above, and *spoken term detection (STD)*.

Most research in speech technology has traditionally been conducted on a small set of well-resourced languages, but the need to extend such technology to many more languages is widely recognized (Patel et al., 2010; Barnard et al., 2010). A variety of approaches for tasks such as corpus development (de Vries et al., 2011) and rapid recognizer bootstrapping (Hughes et al., 2010; Schultz, 2000) have been developed and applied. Spoken Term Detection for under-resourced languages has recently attracted attention, in the form of systems (Hazen et al., 2009; Zhang and Glass, 2009; Muscariello et al., 2011) that employ dynamic time warping to find matches between query terms and the material to be searched. In order to achieve speaker independence, both query and reference speech are represented in terms of *posteriorgrams* – that is, a frame-synchronous series of vectors, each containing estimated posterior probabilities. In Hazen et al. (2009), these are posterior probabilities of the phonetic classes, whereas (Zhang and Glass, 2009; Zhang et al., 2012) employ posterior probabilities of classes determined with unsupervised clustering. Note that the query is itself assumed to exist in spoken form – hence, these methods start from somewhat different assumptions than the conventional STD systems, which assume queries in text form. Spoken input of query terms is characterized as "query by example" in Hazen et al. (2009).

Conventional state-of-the-art STD methods employ general-purpose speech-recognition systems to generate a lattice of word hypotheses (Miller et al., 2007; Chelba et al., 2008). The lattice is then used to create an index of word occurrences within the audio data, and the corresponding confidence scores for each of these detections. During retrieval, it is then simply a matter of finding those word strings that correspond to the query terms. One complication to this conceptually simple approach is that the search terms may not be present in the recognition vocabulary that was used for speech recognition. Hence, alternative representations that can also capture out-of-vocabulary words are required – for example, each word in the recognized lattice can be expressed in terms of its constituent phones. During retrieval, out-of-vocabulary words can then be found by matching their sub-word (for example, phonetic) representation against such phone-based lattices (Wallace et al., 2007). Other sub-word decompositions have also been employed successfully (Szöke, 2010), but in most cases detection of out-of-vocabulary words remains substantially inferior to that of in-vocabulary words.

Although the query by example methods in Hazen et al. (2009), Zhang and Glass (2009), Zhang et al. (2012) achieve promising detection rates, retrieval is significantly more demanding computationally than with index-based approaches (Wallace et al., 2007). In Hazen et al. (2009), it is therefore recommended that such methods be used as a rescoring mechanism for terms retrieved by a cruder (index-based) approach. Also, these methods have to date been assessed on well-resourced languages; hence, issues in their application in real under-resourced environments have not been explored.

The work presented in this paper is different from IARPA's currently active "Babel" program (IARPA, 2011) in two key aspects: we use typically an order of magnitude less data per language and the focus is on language independent approaches, rather than a capability to rapidly bootstrap systems in new languages. SWS-like technologies could also be helpful to the intelligence or military community, for example in quickly changing theaters of operations that span multiple linguistic groups, to develop triage capabilities for intercepted radio communications, or as part of other tactical solutions. Another useful aspect of the described work is that it could be very useful to implement (initial) speech processing capabilities for use by non-speech recognition experts to apply in other research contexts (Kumar et al., 2013). More discussion of related work in ASR can be found in Miller et al. (2007), Shen et al. (2009), Larson et al. (2012), and (Zero resource, 2012).

3. Spoken Web Search at MediaEval

MediaEval is a benchmarking initiative dedicated to evaluating new algorithms for multimedia access and retrieval (MediaEval, 2014). The "Spoken Web Search" (SWS) task was run in 2011 (Rajput and Metze, 2011) and 2012 (Metze et al., 2012), 2013 (MediaEval, 2013), and will again be run in October 2014, following a model in which participants receive labeled development data, before receiving unseen evaluation data a couple of months later, on which they blindly submit results to the organizers for scoring.

Table 1
Development (Dev) and evaluation (Eval) corpora used for the "Spoken Web Search Task" at MediaEval (Metze et al., 2012, 2013). Even though we pose the evaluation as a "Spoken Term Detection" problem, SWS can also be seen as an Information Retrieval (IR) problem, involving "queries" and "documents".

Category	2011 ("Indian")			2012 ("African")		
	# Utts	Total (h)	Avg. (s)	#*Utts*	Total (h)	Avg. (s)
Dev docs	400	2:02:22	18.3	1,580	3:41:52	8.4
Dev queries	64	0:01:19	1.2	100	0:02:22	1.4
Eval docs	200	0:47:04	14.1	1,660	3:52:32	8.4
Eval queries	36	0:00:58	1.6	100	0:02:32	1.5
Total	700	2:51:42	14.7	3,440	7:35:18	7.9

By design, the pilot "Indian" dataset, which was used in the 2011 evaluation (Rajput and Metze, 2011), and was retained as a "progress" set for the 2012 evaluation (Metze et al., 2012), consisted of only 700 utterances in telephony quality (8 kHz/16 bit) from four Indian languages (English, Gujarati, Hindi, Telugu). The data was provided by the Spoken Web team at IBM Research India (Kumar et al., 2007) for research purposes. The "African" dataset replaced the Indian data as primary condition in 2012 to provide more data and variety, while attempting to match the Indian dataset's overall characteristics. It comprises more than 3000 utterances from isiNdebele, Siswati, Tshivenda, and Xitsonga, taken from the LWAZI corpus (Barnard et al., 2009).

For the African data set, query terms ranging from one to three words per term were selected from the development and evaluation sets, in such a way as to produce a range of occurrence frequencies in both these sets. No speaker overlap was allowed between speakers in the development set, the evaluation and/or the query set. As some of the languages are agglutinative, possible search terms that overlap with very frequently occurring words were excluded from the set of queries used. If for example 'asajamile' occurs frequently, then 'ajamile' would be excluded. Asking participants to make this distinction would require them to perform morphological analysis – a task outside the scope of the current challenge. Table 1 lists the characteristics of the SWS datasets.

Languages were represented equally within the respective sets, and language labels were provided only on the development data, in order to represent a realistic scenario. Word-level transcriptions (and corresponding phonetic dictionary entries) were also made available for the training and development sections of the data only. The task therefore required researchers to build a language-independent audio search system so that, given a query, it should be able to find the appropriate audio file(s) and the (approximate) location of a query term within the audio file(s).

Targets were defined by an exact string match of the query term in the reference transcription. A target was "hit", if the system returned a positive detection decision within a temporal window around the reference alignment, while it was "missed" if no positive detection was returned. As explained above, the temporal "cushion" was set to a large value for the "Indian" data, because no exact temporal alignment was available, and utterances were quite short anyway.

By design, performing language identification, followed by standard speech-to-text is not an appropriate approach to SWS, because full-fledged recognizers are typically not available in these languages.

In order to not restrict participation, the use of external resources was permitted, as long as their use had been declared ("open" condition). Systems that were only developed on the provided data are called "restricted". Throughout this paper, we will use the nomenclature shown in Table 2 when discussing and comparing approaches and systems.

3.1. Evaluation and scoring

SWS experiments were scored with a modified version of the NIST 2006 STD evaluation scoring software (Fiscus et al., 2007). The primary evaluation metric was ATWV (Actual Term-Weighted Value), while MTWV (Maximum Term-Weighted Value) was also reported. According to Fiscus et al. (2007), the Term-Weighted value is computed as a function of the miss and false alarm probabilities (P_{miss}, P_{FA}) averaged over all query terms. The "Actual" TWV is obtained by computing the value for a given operating point set by the participant, "Max" TWV is determined by selecting the optimum operating point over all queries.

As no accurate word alignment was available for the "Indian" data, a hypothesized match was considered correct provided that it occured in the correct reference file, by setting wide temporal padding parameters. On the "African"

Table 2
Classification of (primary) SWS submissions: "open" means that external data sources were used, "restricted" means that only the resources provided during that year's evaluation were used. A "symbol-based" system computes similarities at some symbolic level, while a "frame-based" system computes and sums distances over time during decision making.

Approach	Condition			
	2011		2012	
	Open	Restricted	Open	Restricted
Symbol-based	Barnard et al. (2011), Szöke et al. (2011), and Mantena et al. (2011)	–	Abad and Astudillo (2012), Varona et al. (2012), and Buzo et al. (2012)	Szöke et al. (2012)
Frame-based	–	Anguera (2011) and Muscariello and Gravier (2011)	Wang and Lee (2012)	Anguera (2012), Jansen et al. (2012), Joder et al. (2012), and Vavrek et al. (2012)

data, accurate ground truth alignments were available, therefore these parameters were set back to the NIST standard parameters. Additionally, in 2012 a modification was incorporated to weight missed and false alarm detections differently in the final metric. In the default NIST setting, the impact of a false alarm is three orders of magnitude that of a miss. While these settings are adequate for a large data monitoring scenario, they might not be appripriate for a retrieval scenario, like the one proposed in this evaluation. The exact impact of an individual false alarm detection varies with the length of the reference database and the number of query terms used, so that care has to be taken when partitioning or combining data sets or query lists. For SWS 2012, settings in the NIST scoring scripts were modified to ensure that $P_{miss} = P_{FA}$ for an equal number of misses and false alarms in the output.

A patched scoring package was distributed along with the data.

3.2. Overall system architecture

The 2011 pilot evaluation attracted the interest of 5 sites, while 9 teams participated in the 2012 evaluation. Competing systems implemented a wide range of solutions. Nevertheless, all systems follow the same overall architecture, illustrated in Figure 1. The first step is front-end processing which comprises several steps such as silence detection, feature extraction and normalization. The second step corresponds to the actual search where the database is matched against the query using either a symbolic representation or frame-based pattern matching. The search procedure computes a score for each attempted match for the utterances in the database. The last step is the decision making step where actual answers to the query are selected from the search result. This step is often limited to the comparison to a decision threshold, possibly after score normalization.

We review each of these three steps in turn in the next sections: "Front-End" and "Database" in Section 4, Frame-based "Search" in Section 5, Symbolic "Search" in Section 6, and the "Decision" step in Section 7.

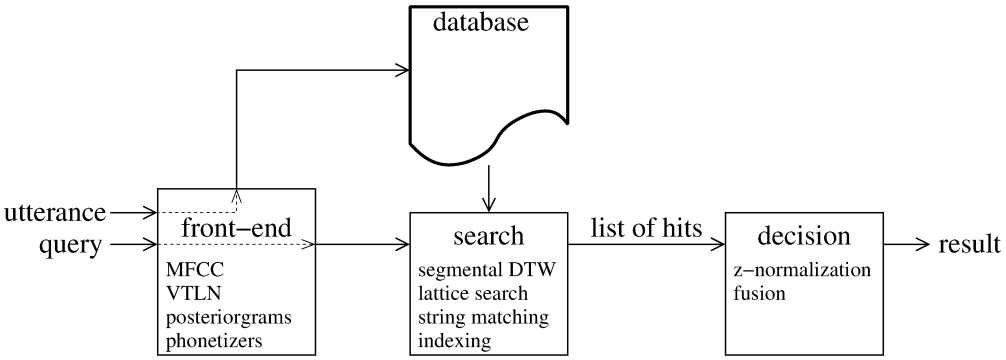

Figure 1. Generic architecture of typical "Spoken Web Search" systems, as implemented by participants.

4. Front-end

In this section, we review such fundamental underlying techniques used for low-resource STD such as feature extraction, feature normalization, voice activity detection and tokenizer development.

4.1. Feature extraction

Most common acoustic features included standard MFCC (Joder et al., 2012; Wang and Lee, 2012; Vavrek et al., 2012; Barnard et al., 2011; Mantena et al., 2011), Bottle-neck (Szöke et al., 2011, 2012), Frequency domain linear prediction (FDLP-S) (Jansen et al., 2012) and PLP features (Abad and Astudillo, 2012). Apart from being used in the development of full-fledged tokenizers (c.f. Section 4.4), these features were also used in direct frame-based comparisons, or as a preliminary step to obtain posteriorgram-based features (Wang and Lee, 2012; Muscariello and Gravier, 2011).

Posteriorgrams were obtained in different manners. One approach was the use of phoneme tokenizers trained using external data. In this case the posterior probabilities of each phoneme were concatenated into a feature vector. As phoneme tokenizers are inherently dependent on a language, other common approaches were to use posterior probabilities from language independent models such as GMMs directly trained on the data (Anguera, 2011, 2012; Muscariello and Gravier, 2011; Wang and Lee, 2012) or using automatically derived acoustic units obtained on the development dataset (Wang and Lee, 2012).

4.2. Feature normalization

Some systems applied methods to normalize the features so as to minimize dependencies to speakers and acoustic conditions, which is especially relevant for frame-based systems. In particular, it was observed that applying cepstral mean and variance normalization improved the matching accuracies (Muscariello and Gravier, 2011; Anguera, 2012; Joder et al., 2012). In addition, two systems showed substantial improvements by applying vocal tract length normalization (VTLN) (Wang and Lee, 2012; Szöke et al., 2012).

4.3. Voice activity detection

Given the spontaneous nature of the acoustic data and the fact that in MediaEval 2012 some of the queries contained multiple words, variable amounts of non-speech were present in the recordings, which always causes a problem for frame-based matching systems and can potentially lead symbol-based systems to obtain wrong decodings. For this reason, several systems proposed and implemented various voice activity detection algorithms (Anguera, 2012; Wang and Lee, 2012; Jansen et al., 2012; Szöke et al., 2012). Examples include unsupervised training of a 2-class speech/non-speech classifier using GMMs on MFCC features (Anguera, 2012), or the use of posteriors from given phone tokenizers (Szöke et al., 2012).

4.4. Tokenizer development

All symbol-based and a few frame-based approaches required the development of a full-fledged phonetic tokenizer, or a set of tokenizers. For symbol-based approaches, the tokenizers produce phone strings or lattices for further analysis; in the frame-based approach, the tokenizer is used to generate posteriorgrams.

The vast majority of systems based phonetic tokenizers on hidden Markov models (Mantena et al., 2011; Barnard et al., 2011; Varona et al., 2012; Szöke et al., 2011), using standard training procedures. Most teams utilized external data in order to optimize the accuracy of their tokenizers, apart from Wang and Lee (2012) where automatically derived phonetic-like units were derived from the development data. This is described in more detail in Section 6.3.

Systems varied quite significantly with regard to the actual number of base units in the model; typically, symbol-based techniques are very sensitive toward this parameter. Base units were in all cases approximations of phones or phone groups, with phone sets mostly reduced, often quite aggressively, for example, from 77 to 28 in the case of Buzo et al. (2012) and 62 to 43, and then to 21 in Barnard et al. (2011). Systems tried to compensate for limited acoustic data (and language mismatch when using external data) by modeling broad phonemic classes rather than detailed phonemes.

This reduction was mostly achieved by merging similar phones based on their IPA identity (Buzo et al., 2012; Varona et al., 2012; Barnard et al., 2011), and in one case, by using articulatory features to combine models (Mantena et al., 2011).

5. Frame-based approaches

Frame-based approaches (often also being referred to as *pattern-based* or *pattern matching* approaches) perform the matching of query audio and the reference data at the acoustic feature frame level. Such matching relies only on the local similarity between frame pairs and posterior time-alignment (allowing for frame insertions and deletions) of pairs in sequence. All frame-based approaches submitted to the evaluation use dynamic programming algorithms inspired by Dynamic Time Warping (DTW) for posterior time alignments, often implementing segmental versions to account for unknown start and end points of matches in the search database. Several variations were proposed, depending on the front-end processing and on the time-alignment algorithm.

When used in MediaEval's frame-based approaches, posteriorgrams turned out to generally outperform raw acoustic features such as MFCCs due to their increased robustness to speaker and acoustic variability.

Many variants of dynamic alignments were used for frame-based search. Standard DTW was considered in some systems, with start and end points determined prior to DTW. In Mantena et al. (2011), a rough acoustic decoding step based on articulatory features is used to find putative matching regions on which DTW is to be applied. A brute-force approach was taken in Szöke et al. (2011) where the DTW similarity is calculated for all the possible combinations of start and end points. As an alternative to standard DTW, segmental variants were used, where start and end points were decided as part of the optimal alignments rather than defined in a pre-processing step. In Muscariello and Gravier (2011), segmental locally normalized DTW (Muscariello et al., 2012) was used to find potential matches of the query. In Anguera (2011), a sub-sequence DTW algorithm (Anguera and Ferrarons, 2013) was used, using either posterior probability features or binary features for efficiency. Another variant of DTW, called "cumulative DTW", was used in Joder et al. (2012) where the usual maximization is replaced by a soft-max rule. Moreover, the pairwise local distances were replaced by step functions resulting from the combination of feature functions where the combination parameters were learned from the data.

Interestingly, a different search strategy was used in Jansen et al. (2012) based on Hough transforms to find near-diagonal lines in a sparse similarity matrix obtained from locality sensitive hashing (LSH) of raw features. The combination of a fast Hough transform and frame indexing (Jansen and Durme, 2012) offers substantial potential in terms of speed and scalability.

Any of the search strategies mentioned above returned a set of putative matching segments which were, in most cases, post-processed to refine the matches before making a decision. In Vavrek et al. (2012), SVM classification was applied to the output of the DTW alignment to decide whether the alignment corresponds to a match or not. In Muscariello and Gravier (2011), an image-based comparison of the query and reference segments represented as self-similarity matrices was performed to increase robustness to speaker and acoustic conditions (Muscariello et al., 2012). A self-similarity matrix is a square, symmetric matrix of distances computed between the individual frames of an utterance. If two sequences are similar, so are the respective self-similarity matrices, which can therefore be used as a distance measure itself. Similarly, putative matches resulting from the Hough transform in Jansen et al. (2012) were further validated using standard DTW.

6. Symbol-based approaches

Symbol-based approaches to the SWS task first convert the query and the content into a symbolic representation, on which the best match is then computed without taking temporal alignment into account. While these symbols can, in principle, be any categorical representation, all SWS submissions used symbol sets that were either based on phones or broad phonemic classes, defined at a coarser granularity than would typically be used in standard speech recognition.

Since the success of symbol-based approaches relies heavily on the accuracy of the tokenizer, it is not surprising that most participants used tokenizers from the "open" category, and used additional resources during system development. Only one participant (Szöke et al., 2012) created a symbol-based system without utilizing any external resources.

Two main approaches were used:

- Acoustic keyword spotting (AKWS), in which a tokenized query string competes with a background and/or filler model during decoding (Abad and Astudillo, 2012; Szöke et al., 2012, 2011).
- String matching, using a form of DTW at the symbol level. Queries were mostly converted to single strings, while content utterances were alternatively represented as strings, *n*-best lists, lattices or confusion networks (Varona et al., 2012; Buzo et al., 2012; Barnard et al., 2011; Mantena et al., 2011).

6.1. Acoustic keyword spotting

Apart from the development of the actual tokenizers (c.f. Section 4.4) approaches differed mainly with regard to the architecture of the AKWS system and normalizations applied (c.f. Section 7).

The system in Szöke et al. (2012) created an HMM for each query, and calculated the log-likelihood ratio between the query and a background/filler model (Schwarz, 2009; Szöke et al., 2005), implementing the background model as a free phone loop without any weighting. In Abad and Astudillo (2012), a hybrid ANN/HMM approach is employed, e.g., the HMM is used to model the speech signal, and the ANN to estimate the posterior phone probabilities. A sliding window was used to process each file, with a uniform 1-g language model formed by the target query and a competing speech background model being used, and the weight of the background model tuned on the development set.

In all cases query pronunciations were obtained by tokenizing the audio automatically (Szöke et al., 2012, 2011), with (Abad and Astudillo, 2012) optimizing on a development set to obtain the correct number of phonemes. All systems used single pronunciations, with some experiments in creating additional variants not producing a win (Abad and Astudillo, 2012). In a separate experiment (in-admissible for the primary condition) the sensitivity of the system was tested toward the accuracy of the derived dictionary, finding that force-aligned transcriptions resulted in a significant TWV increase of about 0.28 (Szöke et al., 2012).

6.2. String matching

All string matching approaches used some form of DTW to match query to content, and differed mainly with regard to the content representation used and the cost function employed during DTW. Apart from one use of an articulatory-inspired phone set (only differentiating between articulation effects rather than individual phones) (Mantena et al., 2011), all systems used fairly standard phone sets, reducing the number of phonemes for increased generalization and improved accuracy given that these had been usually trained with data from other languages.

Basic string-based DTW, with each content utterance also being represented as a single string, was implemented by most participants (Buzo et al., 2012; Barnard et al., 2011; Mantena et al., 2011). Strings are tokenized, DTW is applied with a sliding window and some form of filtering used to remove overlapping queries. In addition, DTW search was also performed within confusion networks (Mangu et al., 2000), with the score weighted according to the alignment of the query with the confusion network (Buzo et al., 2012), and within lattices (Barnard et al., 2011; Varona et al., 2012). Specifically, the Varona et al. (2012) system extracted the N phone decodings with the highest likelihoods from the phone lattice and then converted them to multigrams (Wang et al., 2011). Results were filtered based on rank as well as score, after normalization.

One of the key element of string matching is the cost matrix which holds the cost of substituting one unit (e.g., phone) with another. Cost matrices were either flat (Mantena et al., 2011), linguistically motivated (Barnard et al., 2011), or estimated from development data.

6.3. Cross-language data sharing

While the main objective of the "Spoken Web Search" task is to search languages with very limited training resources, many systems in the MediaEval evaluation found it useful to utilize data resources from other languages.

Systems in the "open" category differed substantially with regard to the amount and type of data that were incorporated from additional sources; an approach that was only used as basis for symbol-based techniques. Only two groups used data of the same language family as the target data, these being Telugu (Mantena et al., 2011) and Hindi (Barnard et al., 2011) in MediaEval 2011. For the rest, data was sourced from various other languages (Romanian,

Czech, English, Hungarian, Levantine, Polish, Russian, Slovak, European and Brazilian Portuguese, European Spanish, American English, etc.), often based on the BUT phone recognizers (Phoneme recognizer, 2009).

External data was incorporated into primary systems by either using the foreign models directly, tokenizing the queries and content using the foreign tokenizers directly (Szöke et al., 2011; Abad and Astudillo, 2012), or by developing tokenizers with foreign data, and then adapting these. Adaptation was performed in different ways. The Buzo et al. (2012) system used MAP adaptation, first low-pass filtering broad-band data, then mapping phones using International Phonetic Alphabet tables or a confusion matrix, tuning the phone error rate on the MediaEval development set. The Barnard et al. (2011) system MAP-adapted the same source data to the four target languages to create four separate tokenizers. The Szöke et al. (2011) system used language-specific Karhunen–Loeve transforms during adaptation.

7. Normalization and decision making

As a result of the search step, a set of possible matches is obtained for every query term. For scoring, a match must include the matching utterance ID and the start-end points where the query is thought to appear, together with a relevance score. Several methods have been proposed in both symbol-based and frame-based approaches to improve the matching results at this point. These include score normalization techniques and fusion of different system outputs.

7.1. Score normalization

The queries that were used in the evaluation have very different acoustic characteristics. On the one hand, their length (before any voice activity detection was applied) ranged from 0.39 s to 4.12 s in the development set, and between 0.38 s and 5.96 s in the evaluation set. On the other hand, some queries consisted of one single word while others had two, or more words. In addition, each phonetic class has different average matching scores: stable parts in vowels and silences have a very good intra-class match, for example, while consonants achieve lower direct matching scores. For these reasons, the distributions of scores for reference matching sequences to each query usually differs quite a bit among queries.

Several methods have been proposed to normalize such scores in order to allow for the application of a single optimal detection threshold. In Wang and Lee (2012), Joder et al. (2012), and Jansen et al. (2012) various flavors of z-norm normalization were applied. In Wang and Lee (2012) the normalization mean and variance for each query in the test was estimated by using development data. On the contrary, Jansen et al. (2012) used the set of possible matches of test queries with the test data to compute such parameters. Similarly, in Joder et al. (2012) the test data was used to find appropriate normalization parameters, although the authors avoid using the 10% best-matching scores to avoid a bias with the actual matches.

A totally different approach was followed by Szöke et al. (2012), where a linear regression model was trained using development data to predict the ideal threshold. Parameters such as the query length, total amount of detected silence in the query, number of phonemes, and so on were used. In all cases, the systems were reported to improve results by using such approaches.

7.2. Intra-system fusion

Several groups submitted the output of a fusion of multiple systems as their primary submission, and reported consistent gains. A multitude of techniques were tried:

Wang and Lee (2012) achieved system combination (or fusion) by averaging the distance matrices computed with different tokenizers before computing DTW (Wang et al., 2013). Individual tokenizers were based on different training data, or target sets. When going from a two-system combination to a seven-system combination, gains reach 0.1 (in terms of ATWV) on the dev data, and 0.15 on the eval data, when going from a five-system combination to the seven-system combination, gains are less than or equal to 0.02.

Abad and Astudillo (2012) explored system combination using “AND”, “OR”, and “MAJORITY” operations on four individual sub-system outputs. Majority voting was found to give the best result, but no performance numbers have been published for the individual sub-systems.

In a different line of thought, pseudo-relevance feedback was used in Wang and Lee (2012), where the top matches obtained from the original query were used to rescore the remaining matches, with the goal to obtain a better score

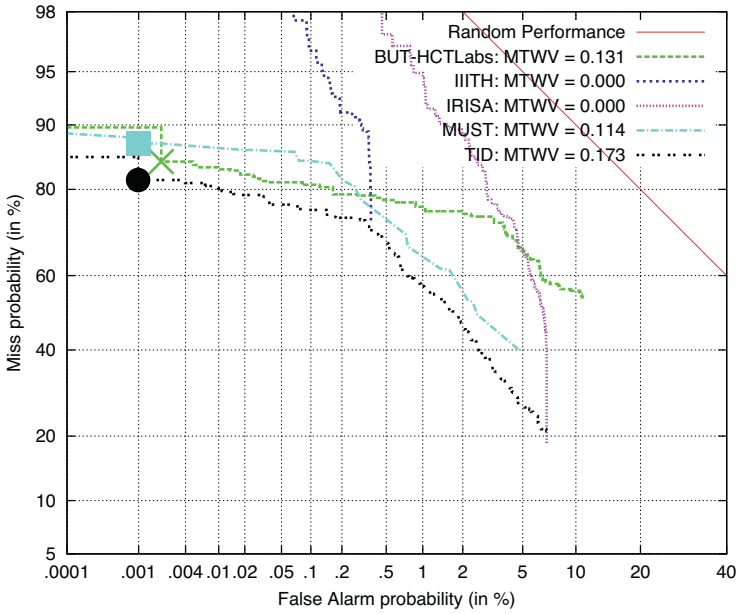

Figure 2. Results (ATWV plot and MTWV) for evaluation terms on the 2011 eval data (Metze et al., 2012). The operating point of the participants' submissions for their respective ATWV is indicated by the marker on the line.

estimation. In Anguera (2012) and Mantena et al. (2011) overlap detection on the resulting matches was used to merge overlapping results.

8. Discussion

Results for all primary and the most relevant contrastive systems are presented in Figures 2 and 3 for the 2011 and 2012 benchmarks respectively. We also report ATWV in Table 3 for the 2012 data. Note that the 2011 and 2012 results were established on two distinct data sets, using different scoring parameters, and are therefore not directly comparable. In spite of this difference in the dataset, it is clear that progress has been made between the two benchmarks, with new

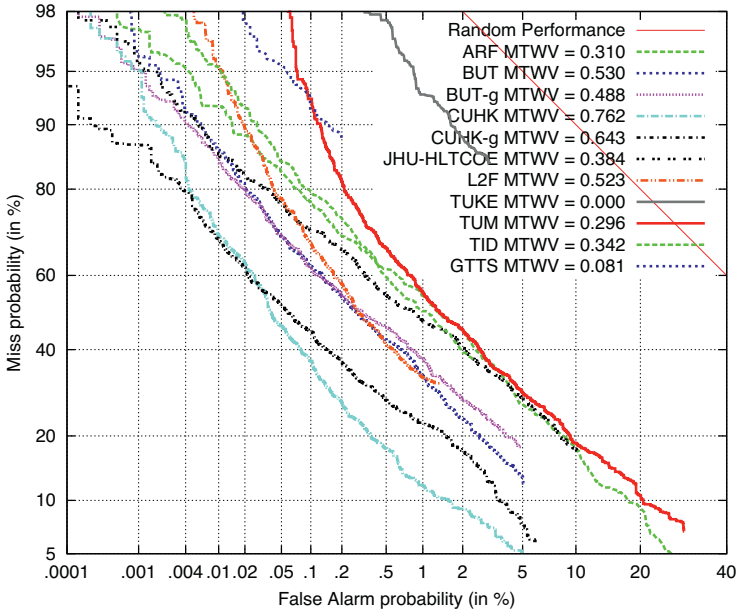

Figure 3. ATWV plots and MTWV results for evaluation terms on the 2012 eval data (Metze et al., 2013). Operating points are not shown for clarity of presentation. Figures 2 and 4 show the difference between well-tuned operating points for the scoring defined on "Indian" and "African" data sets by the organizers.

Table 3
Results (actual TWV) for selected SWS 2012 systems (Metze et al., 2013).

System	Type	Dev	Eval	See
cuhk_phnrecgmmasm_p-fusionprf (CUHK)	Open	0.782	0.743	Wang and Lee (2012)
cuhk_spch_p-gmmasmprf (CUHK-g)	Restricted	0.678	0.635	Wang and Lee (2012)
l2f_12_spch_p-phonetic4_fusion_mv	Open	0.531	0.520	Abad and Astudillo (2012)
but_spch_p-akws-devterms (BUT)	Open	0.488	0.492	Szöke et al. (2012)
but_spch_g-DTW-devterms (BUT-g)	Open	0.443	0.448	Szöke et al. (2012)
jhu_all_spch_p-rails (JHU-HLTCOE)	Restricted	0.381	0.369	Jansen et al. (2012)
tid_sws2012_IRDTW	Restricted	0.387	0.330	Anguera (2012)
tum_spch_p-cdtw	Restricted	0.263	0.290	Joder et al. (2012)
arf_spch_p-asrDTWAlign_w15_a08_b04	Open	0.411	0.245	Buzo et al. (2012)
gtts_spch_p-phone_lattice	Open	0.098	0.081	Varona et al. (2012)
tuke_spch_p-dtwsvm	Restricted	0	0	Vavrek et al. (2012)

pre-processing, tokenization and normalization techniques appearing in 2012. In the following, we first review and comment on evaluation results, before presenting fusion experiments and initial analysis of how language properties impact detection performance.

While in 2011 few systems implemented frame-based approaches using pattern matching techniques, such approaches were implemented in the majority of the 2012 submissions. Moreover, in both benchmarks, the best results were obtained by template matching systems. Frame-based, template matching techniques are gaining interest in the community and can achieve the same performance as symbol-based approaches on the "operating point" chosen for MediaEval with respect to amount and kind of data. The BUT experiments (Szöke et al., 2012) show that not having a lexicon available greatly impacts the performance of AKWS systems, which are limited to extracting information from one, isolated query only in this scenario. Computation may be an issue for frame-based systems, but techniques have been developed to search even large amounts of data efficiently (Jansen et al., 2012). The best systems tend to combine multiple representations and techniques, achieving significant gains in the process; we speculate that SWS could be an interesting, low-complexity test-bed to develop complementary data representations, matching techniques, and search approaches. It is however interesting to note that under the SWS evaluation conditions, the zero-knowledge ("restricted") approaches performed quite similarly to "open" (typically model-based) approaches, which typically rely on the availability of matching data from other languages. On the 2012 evaluation, the difference in ATWV is about 0.1 for the two CUHK systems (Wang and Lee, 2012) and 0.05 for the two BUT systems (Szöke et al., 2012).

As a first observation, the larger set of participants also resulted in a variety of new signal-processing techniques being introduced or old techniques being re-introduced, such as Vocal Tract Length Normalization (VTLN), which boosted performance significantly for the CUHK system (Wang and Lee, 2012), although no detailed analysis has been performed on this isolated aspect. Similarly, cepstral mean subtraction and variance normalization have been successfully applied to most pattern matching systems. 2011 results had also alerted participants to the importance of silence segmentation for this type of task, as silence segments should not by counted in any distance or matching function. With model-based tokenizers increasingly exploiting techniques from speech and speaker recognition, we expext that normalization techniques, such as VTLN, SAT or factor analysis will be adapted to the spoken Web search task in the near future.

Secondly, efforts have been devoted between 2011 and 2012 to the selection of suitable acoustic units in the tokenizers. A greater variety of languages were used, and data-driven units have successfully been combined with phonetic units in the "open" condition. The best performing system in 2012 is a frame-based system linearly combining similarity matrices obtained in different ways, either in a restricted mode (GMM, self-trained acoustic units) or in an open mode (phone models from various languages). Intermediate results reported in Wang and Lee (2012) show that the combination of phone models from 5 different languages (cz, hu, ru, ma, en) gives better results than the combination of GMM and ASM posteriorgrams (ATWV 0.72 vs. 0.59 on the 2012 evaluation data), while using all 7 tokenizers outperforms both settings (ATWV of 0.74). A more detailed description of this system is available in Wang et al. (2013). The combination of distance matrices obtained from posteriorgrams with different tokenizers could also have contributed to an increased robustness to speaker variability.

Table 4
TWV values on the 2012 data set for the different African languages.

Language	ATWV		MTWV	
	Open	Restricted	Open	Restricted
IsiNdebele	0.609	0.512	0.717	0.593
Siswati	0.644	0.583	0.782	0.709
Tshivenda	0.718	0.604	0.718	0.612
Xitsonga	0.650	0.559	0.650	0.613
All	0.698	0.586	0.763	0.658

Finally, score normalization techniques derived from speaker verification were introduced by most participants in 2012 so as to normalize scores on a per-query basis, to good effect.

Most participants chose a z-norm-like scheme, normalizing scores with respect to the query, which is fairly easy to implement and reported benefits of such a normalization in the working note papers (MediaEval Benchmark, 2011, 2012). Various flavors of z-norm were used (under different names), but the absence of contrastive results does not allow to vouch for one or another. At the 2012 evaluation workshop, participants expressed an interest in exploring t-norm techniques, at normalizing the decision score with respect to the document in which the query is searched for, thereby integrating more insight from speaker recognition work. While such techniques would undoubtedly improve all approaches, their implementation is both algorithmically complex and computationally intensive and was not considered so far.

In light of these elements of analysis, we believe that speaker normalization techniques (VTLN, mean and variance normalization), along with the combination of multiple distance matrices obtained from different posteriorgrams are the key features explaining the success of the CUHK system (Wang and Lee, 2012; Wang et al., 2013) over other participants.

Focusing on frame-based systems, results from TID (Anguera, 2012; Mantena and Anguera, 2013) and JHU-HLTCOE (Jansen et al., 2012) are of particular interest: both systems implement indexing techniques for efficient pattern matching, where other frame-based approaches rely on segmental variants of the dynamic time warping algorithm. Such variants are typically computationally expensive and severely limit the scalability of frame-based approaches. On the contrary, indexing techniques enable the efficient computation of a sparse similarity matrix, whose sparsity in turn enables fast matching. While the two approaches exploiting indexing techniques do not outperform others, they exhibit a fair performance level while being scalable. These two systems thus clearly demonstrate that combining indexing techniques such as LSH or efficient approximate nearest neighbor search with pattern matching is a valid research trend for fast and scalable language-independent spoken term detection. During the evaluation, participants were encouraged to measure and report the computational requirements of their approaches; however, the wide variety of resources used make a fair comparison between systems difficult at this point.

8.1. Multi-site fusion

Similarly to the well-known ROVER approach to combining multiple speech-to-text systems (Fiscus, 1997), the results of multiple, independent STD systems can also be combined. Several methods can be employed to generate appropriate combination weights, such as maximum entropy or linear regression (Norouzian et al., 2012). For symbolic-based systems, the combination of pronunciation dictionaries, which were generated using different approaches is viable (Wang, 2010). This approach however is not suitable for dictionary-less systems, such as the "frame-based" approaches discussed here. In any case, scores or posteriors need to be normalized suitably across systems for successful normalization. Several well-performing participants also performed system combination using score averaging (Wang and Lee, 2012) or voting (Abad and Astudillo, 2012).

To combine the output of all MediaEval submissions, we employed the CombMNZ algorithm (Fox and Shaw, 1994). CombMNZ is a general data-fusion technique, which still requires some score normalization, as previously discussed. In the authors' own work on IARPA's Babel program (IARPA, 2011), this algorithm provided almost always the best performance across a wide range of conditions, and was certainly the most robust fusion technique. Table 4 shows the

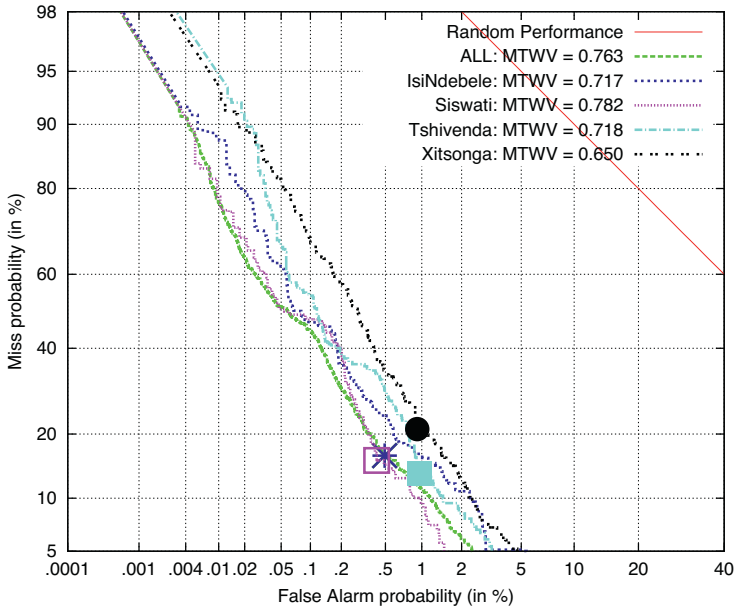

Figure 4. Results (ATWV plots and MTWV) on the 2012 evaluation data on a per language basis for a combination of open systems.

results of this "fused" system for the 2012 evaluation data. Figures 4 and 5 show that the corresponding curves are somewhat more "well-behaved", even if the Maximum TWV could not be improved, so there is some benefit from system fusion even in this case.

Given the large advantage that the CUHK systems (Wang and Lee, 2012) enjoyed over the other participants in the 2012 evaluation, the organizers were not able to improve the performance significantly by merging multiple system outputs (ranked lists), but the combination of several other systems provided good gains, for both the "open" and "restricted" cases. It may be possible to achieve gains by using a matrix combination technique as described in Wang et al. (2013).

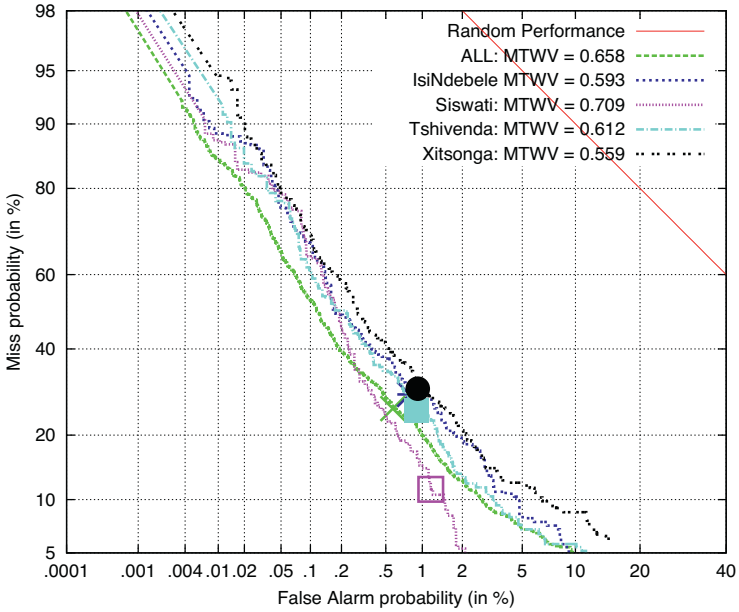

Figure 5. Results (ATWV plots and MTWV) on evaluation data on a per language basis for a combination of restricted systems.

8.2. Influence of language

Figures 4 and 5 provide a breakdown of the 2012 evaluation results for the four languages represented in the database. Results are provided for a fusion of the best open resources systems and a fusion of the best restricted systems, open systems performing slightly better than restricted ones. In both cases, Xitsonga appears more difficult than the other languages while Siswati yields the best performance. We believe that these results can be partially explained by the average word length which differs between languages, where longer average word leads to better results. Regardless of the word length consideration, one can expect that languages most similar to Germanic languages, i.e. Tshivenda and Xitsonga, benefit most from the open condition (Zulu et al., 2008). However, this linguistically motivated expectation was not met (see Table 4), probably because of the influence of stronger non-linguistic factors such as word length and the randomness due to the choice of the queries.

9. Conclusion and outlook

The capability to detect spoken terms or recognize keywords in low or zero resource settings is an important capability which could boost the use of speech technology in developing regions significantly. When there are neither experts, who could develop speech recognition systems and maintain the infrastructure required for designing speech dialogs and indexing audio content, nor databases on which speech interfaces could be developed, zero resource spoken term detection as presented here could provide a "winning combination". In this paper we presented the main findings of the "Spoken Web Search" task within the MediaEval evaluation benchmark, as run in 2011 and 2012. We believe the results achieved in the evaluations show that these techniques could be applied in practical settings, even though user tests are certainly needed to determine the overall performance, and an acceptable ratio of false alarms and missed detections for a given application. It is interesting to note that the very diverse systems presented, which cover a wide range of possible approaches, could achieve very similar results, and future work should include more evaluation criteria, such as amount of external data used, processing time(s), etc., which were deliberately left unrestricted in this evaluation, to encourage participation.

With respect to the amount of actual data available, the SWS task is much harder than the research goals proposed by for example IARPA's Babel (IARPA, 2011) program, where up to 100 h of transcribed data per language are available, and the language of a test query is known. The SWS task is targeted primarily at communities that currently do not have access to the Internet at all. Many target users have low literacy skills, and many speak in languages or dialects for which fully developed speech recognition systems will not exist even for years to come. We hope that the recent surge of activity in zero resource approaches (see e.g. Zero resource, 2012; Jansen et al., 2013) will result in further progress, which will advance the state of the art in spoken term detection and document retrieval significantly, specifically when large data sets and databases are not available.

Acknowledgments

The authors would like to acknowledge the MediaEval Multimedia Benchmark (MediaEval Benchmark, 2012). We especially thank Martha Larson from TU Delft for organizing this event, and the participants for their hard work on this evaluation. The "Spoken Web Search" task was originally proposed by researchers from IBM India (Agarwal et al., 2010). North-West University and IBM Research India collected and provided the audio data and references used in these evaluations.

References

Abad, A., Astudillo, R.F., 2012. The L2F spoken web search system for MediaEval 2012. In: Proc. MediaEval 2012, http://www.multimediaeval.org/mediaeval2012/; http://ceur-ws.org/Vol-927/

Agarwal, S., Kumar, A., Nanavati, A.A., Rajput, N., 2009. Content creation and dissemination by-and-for users in rural areas. In: Proc. Intl. Conf. Information and Communication Technologies and Development (ICTD), Doha, Qatar.

Agarwal, S., Dhanesha, K., Jain, A., Kumar, A., Menon, S., Nanavati, A., Rajput, N., Srivastava, K., Srivastava, S., 2010. Organizational, social and operational implications in delivering ICT solutions: a telecom web case-study. In: Proc. ICTD, London, UK.

Agarwal, S., Jain, A., Kumar, A., Rajput, N.,2010. The world wide telecom web browser. In: Proc. First ACM Symposium on Computing for Development. ACM, London, UK.

Anguera, X., Ferrarons, M., 2013. Memory efficient subsequence DTW for query-by-example spoken term detection. In: ICME: International Conference on Multimedia and Expo, San Jose, CA, USA.

Anguera, X., 2012. Telefonica system for the spoken web search task at MediaEval 2011. In: Proc. MediaEval 2011, http://www.multimediaeval.org/mediaeval2011/; http://ceur-ws.org/Vol-807/

Anguera, X., 2012. Telefonica research system for the spoken web search task at MediaEval 2012. In: Proc. MediaEval 2012, http://www.multimediaeval.org/mediaeval2012/; http://ceur-ws.org/Vol-927/

Barnard, E., Davel, M.H., van Heerden, C.,2009. ASR corpus design for resource-scarce languages. In: Proc. INTERSPEECH. ISCA, Brighton, UK, pp. 2847–2850.

Barnard, E., van Schalkwyk, J., van Heerden, C., Moreno, P.J.,2010. Voice search for development. In: Proc. INTERSPEECH. ISCA, Makuhari, Japan, pp. 282–285.

Barnard, E., Davel, M.H., van Heerden, C., Kleynhans, N., Bali, K., 2011. Phone recognition for spoken web search. In: Proc. MediaEval, 2011, http://www.multimediaeval.org/mediaeval2011/; http://ceur-ws.org/Vol-807/

Buzo, A., Cucu, H., Safta, M., Ionescu, B., Burileanu, C., 2012. ARF@MediaEval 2012: a Romanian ASR-based approach to spoken term detection. In: Proc. MediaEval 2012, http://www.multimediaeval.org/mediaeval2012/; http://ceur-ws.org/Vol-927/

Chelba, C., Hazen, T.J., Saraçlar, M., 2008. Retrieval and browsing of spoken content. IEEE Sign. Process. Mag. 25 (3), 39–49.

de Vries, N.J., Badenhorst, J., Davel, M.H., Barnard, E., de Waal, A.,2011. Woefzela – an open-source platform for ASR data collection in the developing world. In: Proc. INTERSPEECH. ISCA, Florence, Italy, pp. 3177–3180.

Diao, M., Mukherjea, S., Rajput, N., Srivastava, K., 2010. Faceted search and browsing of audio content on spoken web. In: CIKM '10: Proceedings of the nineteenth international conference on Information and knowledge management, Toronto, Canada.

Education for all global monitoring report – reaching the marginalized, 2010. http://unesdoc.unesco.org/images/0018/001866/186606E.pdf. Last accessed: March 1, 2014.

Fiscus, J., Ajot, J., Garofolo, J., Doddington, G., 2007. Results of the 2006 spoken term detection evaluation. In: Proc. SSCS, Amsterdam, Netherlands.

Fiscus, J.,1997. A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER). In: Proc. Automatic Speech Recognition and Understanding Workshop. IEEE, Santa Barbara, CA, USA, pp. 347–354.

Fox, E.A., Shaw, J.A., 1994. Combination of multiple searches. In: Proc. 2nd Text REtrieval Conference (TREC-2), Gaithersburg, MD, USA, pp. 243–252.

Hazen, T.J., Shen, W., White, C.,2009. Query-by-example spoken term detection using phonetic posteriorgram templates. In: Proc. ASRU. IEEE, Merano, Italy.

Hughes, T., Nakajima, K., Ha, L., Vasu, A., Moreno, P., LeBeau, M., 2010. Building transcribed speech corpora quickly and cheaply for many languages. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010), Makuhari, Japan, pp. 1914–1917.

Intelligence Advanced Research Projects Activity, 2011. IARPA-BAA-11-02, http://www.iarpa.gov/Programs/ia/Babel/babel.html. Last accessed: March 1, 2014.

Internet Usage World-Wide by Country, 2010. http://www.infoplease.com/ipa/A0933606.html. Last accessed: March 1, 2014.

Jansen, A., Durme, B.V.,2012. Indexing raw acoustic features for scalable zero resource search. In: Proc. INTERSPEECH. ISCA, Portland, OR, USA.

Jansen, A., van Durme, B., Clark, P., 2012. The JHU-HLTCOE spoken web search system for MediaEval 2012. In: Proc. MediaEval 2012, http://www.multimediaeval.org/mediaeval2012/; http://ceur-ws.org/Vol-927/

Jansen, A., Dupoux, E., Goldwater, S., Johnson, M., Khudanpur, S., Church, K., Feldman, N., Hermansky, H., Metze, F., Rose, R., Seltzer, M., Clark, P., McGraw, I., Varadarajan, B., Bennett, E., Borschinger, B., Chiu, J., Dunbar, E., Fourtassi, A., Harwath, D., Lee, C.-Y., Levin, K., Norouzian, A., Peddinti, V., Richardson, R., Schatz, T., Thomas, S.,2013. A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition. In: Proc. ICASSP. IEEE, Vancouver, BC, Canada.

Joder, C., Weninger, F., Wöllmer, M., Schuller, B., 2012. The TUM cumulative DTW approach for the MediaEval 2012 spoken web search task. In: Proc. MediaEval 2012, http://www.multimediaeval.org/mediaeval2012/; http://ceur-ws.org/Vol-927/

Kotkar, P., Thies, W., Amarasinghe, S., 2008. An audio Wiki for publishing user-generated content in the developing world. In: HCI for Community and International Development (Workshop at CHI), Florence, Italy.

Kumar, A., Rajput, N., Chakraborty, D., Agarwal, S., Nanavati, A.A., 2007. WWTW: a world wide telecom web for developing regions. In: ACM SIGCOMM Workshop on Networked Systems For Developing Regions, Kyoto, Japan.

Kumar, A., Rajput, N., Agarwal, S., Chakraborty, D., Nanavati, A.A., 2008. Organizing the unorganized – employing it to empower the under-privileged. In: Proceedings of the World Wide Web, Beijing, China.

Kumar, A., Metze, F., Wang, W., Kam, M.,2013. Formalizing expert knowledge for developing accurate speech recognizers. In: Proc. INTERSPEECH. ISCA, Lyon, France.

Larson, M., de Jong, F., Kraaij, W., Renals, S., 2012. Introduction to special issue on searching speech. ACM Trans. Inform. Systems 30 (3).

Mangu, L., Brill, E., Stolcke, A., 2000. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. Comput. Speech Language 14 (4), 373–400.

Mantena, G., Anguera, X.,2013. Speed improvements to information retrieval-based dynamic time warping using hierarchical k-means clustering. In: Proc. ICASSP. IEEE, Vancouver, Canada.

Mantena, G.V., Babu, B., Prahallad, K., 2011. SWS task: articulatory phonetic units and sliding DTW. In: Proc. MediaEval 2011, http://www.multimediaeval.org/mediaeval2011/; http://ceur-ws.org/Vol-807/

MediaEval Benchmark, MediaEval 2011 Workshop, http://www.multimediaeval.org/mediaeval2011/; http://ceur-ws.org/Vol-807/

MediaEval Benchmark, MediaEval 2012 Workshop, http://www.multimediaeval.org/mediaeval2012/; http://ceur-ws.org/Vol-927/
MediaEval Benchmark, MediaEval 2013 Workshop, http://www.multimediaeval.org/mediaeval2013/; http://ceur-ws.org/Vol-1043/
MediaEval Benchmark, 2014. http://www.multimediaeval.org/
Metze, F., Rajput, N., Anguera, X., Davel, M., Gravier, G., van Heerden, C., Mantena, G.V., Muscariello, A., Prahallad, K., Szöke, I., Tejedor, J.,2012. The spoken web search task at MediaEval 2011. In: Proc. ICASSP. IEEE, Kyoto, Japan.
Metze, F., Barnard, E., Davel, M., van Heerden, C., Anguera, X., Gravier, G., Rajput, N., 2012. The spoken web search task. In: Proc. MediaEval Workshop, http://www.multimediaeval.org/mediaeval2012/; http://www.multimediaeval.org/mediaeval2012/sws2012/
Metze, F., Anguera, X., Barnard, E., Davel, M., Gravier, G.,2013. The spoken web search task at MediaEval 2012. In: Proc. ICASSP. IEEE, Vancouver, BC, Canada.
Miller, D.R.H., Kleber, M., Kao, C.-L., Kimball, O., Colthurst, T., Lowe, S.A., Schwartz, R.M., Gish, H.,2007. Rapid and accurate spoken term detection. In: Proc. INTERSPEECH. ISCA, Antwerpen, Belgium.
Muscariello, A., Gravier, G., 2012. IRISA MediaEval 2011 spoken web search system. In: Proc. MediaEval 2011, http://www.multimediaeval.org/mediaeval2011/; http://ceur-ws.org/Vol-807/
Muscariello, A., Gravier, G., Bimbot, F.,2011. A zero-resource system for audio-only spoken term detection using a combination of pattern matching techniques. In: Proc. INTERSPEECH. ISCA, Florence, Italy.
Muscariello, A., Gravier, G., Bimbot, F., 2012. Unsupervised motif acquisition in speech via seeded discovery and template matching combination. IEEE Trans. Audio Speech Language 20 (7), 2031–2044.
Norouzian, A., Jansen, A., Rose, R., Thomas, S.,2012. Exploiting discriminative point process models for spoken term detection. In: Proc. INTERSPEECH. ISCA, Portland, OR, USA.
Patel, N., Chittamuru, D., Jain, A., Dave, P., Parikh, T.S.,2010. Avaaj Otalo: a field study of an interactive voice forum for small farmers in rural India. In: CHI '10: Proceedings of the 28th International Conference on Human Factors in Computing Systems. ACM, Atlanta, GA, USA, pp. 733–742.
Phoneme recognizer based on long temporal context, 2009. http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context. Last accessed: March 1, 2014.
Rajput, N., Metze, F., 2011. Spoken web search. In: Proc. MediaEval 2011, http://www.multimediaeval.org/mediaeval2011/; http://ceur-ws.org/Vol-807/
Raza, A.A., Haq, F.U., Tariq, Z., Pervaiz, M., Razaq, S., Saif, U., Rosenfeld, R.,2013. Job opportunities through entertainment: virally spread speech-based services for low-literate users. In: Proc. CHI. ACM, Paris, France.
Schultz, T., 2000. Multilinguale Spracherkennung: Kombination akustischer Modelle zur Portierung auf neue Sprachen. Universität Karlsruhe, Institut für Logik, Komplexität und Deduktionssysteme (Ph.D. thesis).
Schwarz, P., 2009. Phoneme recognition based on long temporal context. Faculty of Information Technology, Brno University of Technology (BUT) (Ph.D. thesis). http://www.fit.vutbr.cz/research/view_pub.php?id=9132
Shen, W., White, C., Hazen, T.J.,2009. A comparison of query-by-example methods for spoken term detection. In: Proc. INTERSPEECH. ISCA, Brighton, UK.
Sherwani, J., Ali, N., Mirza, S., Fatma, A., Memon, Y., Karim, M., Tongia, R., Rosenfeld, R., 2007. Healthline: speech-based access to health information by low-literate users. In: Proc. IEEE/ACM Int'l Conference on Information and Communication Technologies and Development, Bangalore, India.
Szöke, I., Schwarz, P., Matějka, P., Burget, L., Karafiát, M., Černocký, J., 2005. Phoneme based acoustics keyword spotting in informal continuous speech. LNAI 3658, 302–309, http://www.fit.vutbr.cz/research/view_pub.php?id=7882
Szöke, I., Tejedor, J., Fapšo, M., Colás, J., 2011. BUT-HCTLab approaches for spoken web search. In: Proc. MediaEval 2011, http://www.multimediaeval.org/mediaeval2011/; http://ceur-ws.org/Vol-807/
Szöke, I., Fapšo, M., Veselý, K., 2012. BUT 2012 approaches for spoken web search – MediaEval 2012. In: Proc. MediaEval 2012, http://www.multimediaeval.org/mediaeval2012/; http://ceur-ws.org/Vol-927/
Szöke, I., 2010. Hybrid word-subword spoken term detection. Faculty of Information Technology BUT (Ph.D. thesis). http://www.fit.vutbr.cz/research/view_pub.php?id=9375
Varona, A., Penagarikano, M., Rodriguez-Fuentes, L.J., Bordel, G., Diez, M., 2012. GTTS system for the spoken web search task at MediaEval 2012. In: Proc. MediaEval 2012, http://www.multimediaeval.org/mediaeval2012/; http://ceur-ws.org/Vol-927/
Vavrek, J., Pleva, M., Juhár, J., 2012. TUKE MediaEval 2012: spoken web search using DTW and unsupervised SVM. In: Proc. MediaEval 2012, http://www.multimediaeval.org/mediaeval2012/; http://ceur-ws.org/Vol-927/
Wallace, R.G., Vogt, R.J., Sridharan, S.,2007. A phonetic search approach to the 2006 NIST spoken term detection evaluation. In: Proc. INTERSPEECH. ISCA, Antwerpen, Belgium.
Wang, H., Lee, T., 2012. CUHK system for the spoken web search task at MediaEval 2012. In: Proc. MediaEval 2012, http://www.multimediaeval.org/mediaeval2012/; http://ceur-ws.org/Vol-927/
Wang, D., King, S., Frankel, J., 2011. Stochastic pronunciation modelling for out-of-vocabulary spoken term detection. IEEE Trans. Audio, Speech, Language Process. 9 (4), http://homepages.inf.ed.ac.uk/v1dwang2/public/tools/index.html
Wang, H., Lee, T., Leung, C.-C., Ma, B., Li, H.,2013. Using parallel tokenizers with DTW matrix combination for low-resource spoken term detection. In: Proc. ICASSP. IEEE, Vancouver, Canada.
Wang, D., 2010. Out-of-vocabulary spoken term detection. University of Edinburgh (Ph.D. thesis).
Zero resource speech technologies and models of early language acquisition, 2012. http://www.clsp.jhu.edu/workshops/archive/ws-12/groups/mini-workshop/Last accessed: March 1, 2014.
Zhang, Y., Glass, J., 2009. Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams. In: Proc. ASRU, IEEE, Merano, Italy.

Zhang, Y., Salakhutdinov, R., Chang, H.-A., Glass, J., 2012. Resource configurable spoken query detection using deep Boltzmann machines. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 5161–5164.

Zulu, P.N., Botha, G., Barnard, E., 2008. Orthographic measures of language distances between the official south African languages. Literator 29 (1), 1–20.