QUERY-BY-EXAMPLE SPOKEN TERM DETECTION EVALUATION ON LOW-RESOURCE LANGUAGES

Xavier Anguera[1], *Luis J. Rodriguez-Fuentes*[2], *Igor Szöke*[3*], *Andi Buzo*[4], *Florian Metze*[5], *Mikel Penagarikano*[2]

[1]Telefonica Research (Barcelona, Spain)
[2]University of the Basque Country UPV/EHU (Leioa, Spain)
[3]Brno University of Technology (Brno, Czech Republic)
[4]University Politehnica of Bucharest (Bucharest, Romania)
[5]Carnegie Mellon University (Pittsburgh, PA, USA)

xanguera@tid.es, luisjavier.rodriguez@ehu.es, szoke@fit.vutbr.cz, buzo.andi@gmail.com, fmetze@cs.cmu.edu, mikel.penagarikano@ehu.es

ABSTRACT

As part of the MediaEval 2013 benchmark evaluation campaign, the objective of the Spoken Web Search (SWS) task was to perform Query-by-Example Spoken Term Detection (QbE-STD), using spoken queries to retrieve matching segments in a set of audio files. As in previous editions, the SWS 2013 evaluation focused on the development of technology specifically designed to perform speech search in a low-resource setting. In this paper, we first describe the main features of past SWS evaluations and then focus on the 2013 SWS task, in which a special effort was made to prepare a challenging database, including speech in 9 different languages with diverse environment and channel conditions. The main novelties of the submitted systems are reviewed and performance figures are then presented and discussed, demonstrating the feasibility of the proposed task, even under such challenging conditions. Finally, the fusion of the 10 top-performing systems is analyzed. The best fusion provides a 30% relative improvement over the best single system in the evaluation, which proves that a variety of approaches can be effectively combined to bring complementary information in the search for queries.

Index Terms— benchmark evaluation, low-resource languages, query-by-example spoken term detection

1. INTRODUCTION

The MediaEval benchmark evaluation proposes every year since 2010 a set of tasks on multimedia analysis. Since 2011 the task coined as Spoken Web Search (SWS) has been proposed to participants. This task involves searching for audio content, within audio content, using an audio query. The main difference of this evaluation with regard to the Spoken Term Detection (STD) task conducted by NIST in 2006 [10] and, more recently, the OpenKWS13 evaluation [20], is that participants are not given a textual query, but instead one or more spoken examples of a query. In general, such examples are spoken by different speakers than those appearing in the search repository and under different environment/channel conditions. Besides, SWS evaluations are *multilingual*, whereas NIST STD evaluations focus on a single language, which strongly determines the kind of approaches that can be effectively applied in both cases. In fact, the speech datasets used in SWS evaluations involve languages for which little resources (or no resources at all) are available to train a supervised system, which makes the task specially challenging. This means that standard Speech-To-Text (STT) or Acoustic Key-Word Spotting (AKWS) systems are usually not available on these languages and thus adaptation algorithms or purely zero-resource approaches have to be employed.

SWS evaluations aim at pushing the limits of what can be potentially done with languages or dialects that do not usually get the attention of commercial systems. This effort aligns with recent interest in the community to develop algorithms to allow for the easy and robust development of speech technology for any language, in particular for low-resource (minority) languages. Since minority languages do not usually have enough active speakers to justify a strong investment in developing full speech recognition systems, any speech technology that can be adapted to them can make a big difference. SWS evaluations provide a baseline that allows groups to do research on the language-independent search of real-world speech data, with a special focus on low-resource languages. SWS evaluations also provide a forum to test and discuss original research ideas and a suitable workbench for young researchers aiming to get started on speech technologies.

The name of the task is owned to the initial suggestion by IBM Research India, which in 2011 provided the datasets for the first SWS evaluation [16], containing around 3 hours of spontaneous telephone voice messages in 4 languages spoken in India (Indian English, Gujarati, Hindi and Telugu), with equal amounts of data for each language. This first SWS edition was also the first attempt to explore how current speech technologies could cope with difficult acoustic conditions and languages for which limited resources were available to train standard supervised systems. The database consisted on two subsets (for development and evaluation, respectively), each including specific sets of spoken queries and search utterances. Participation amounted to 5 teams, with systems based on acoustic pattern matching and AKWS approaches. For each given query, the submitted systems should provide the list of search utterances where one or more instances of that query had been detected.

A different dataset was used for the SWS 2012 evaluation [18], including speech in 4 African languages (isiNdebele, Siswati, Tshivenda and Xitsonga), extracted from the LWAZI corpus [7]. Like in 2011, two subsets were created, each one with around 4 hours of searchable utterances and 100 queries. Participants had to return

*Igor Szöke was supported by the Czech Science Foundation, under postdoctoral project No. GPP202/12/P567.

the exact locations within the utterances where each of the queries appeared. Some queries were composed of two words with an undefined amount of silence between them, which raised the importance of building systems able to allow such gaps within query matchings (e.g. by applying speech activity detectors to filter them out). Overall, 9 teams participated and submitted results to the evaluation. For a comprehensive analysis of the techniques proposed in the first two SWS evaluations, see [19].

The rest of the paper is organized as follows. Sections 2 and 3 describe the task and the database used in SWS 2013, respectively. Section 4 reviews the metrics used to measure system performance, including the *normalized cross entropy* metric, C_{nxe}, introduced for the first time in this evaluation. Section 5 highlights the main novelties in the algorithmic approaches proposed by participants and presents a thorough analysis of the obtained results. Section 6 presents a post-evaluation study where the top 10 best performing primary systems were fused together (at the score level) to obtain remarkable performance improvements. Finally, in Section 7 conclusions are drawn and some ground is set for future evaluations.

2. THE SWS 2013 EVALUATION SETUP

The SWS2013 evaluation had three important steps: the release of development data to participants, the release of evaluation queries and the deadline for systems output submission. The development data release included the set of development queries and a set of speech utterances for them to be searched on. Unlike in previous years, a single set of utterances was used both for development and for evaluation. This allowed participants to work with a large set of audio files in which two sets of queries had to be searched for. According to general evaluation best practices, this could be seen as a problem, since participants could try to adapt their systems to the data. However, the mixture of languages and acoustic conditions in the search repository is so large that trying to adapt a system to those conditions was not only acceptable but an interesting issue to do research on. Given that utterances in the search repository were shuffled and no side information was provided to participants regarding the spoken language or the acoustic condition for each file, any possible form of adaptation would have to rely on unsupervised algorithms, thereby introducing an interesting line of research.

With the release of the development data, a ground-truth file was also delivered to participants. This file indicated where each query appeared in the search utterances. Each query was referenced using a general query identifier, which did not disclose to participants its transcription. Neither the transcription nor other information was given about regions in the utterances where no query was present. This was done in order not to disclose any information about the language spoken in each utterance or its contents.

Besides providing a single spoken example for every query, additional examples were also collected for two of the languages (10 examples per query for Czech and 3 examples per query for Basque). These were clearly marked in the query set with the identifier that these belonged to, and were given to participants as an optional task, which aimed at analyzing the effect of enhancing their basic systems when multiple examples per query were available. In most cases, these additional examples were not uttered by the same person. When using them, participants did not know whether they all came from the same or from different languages.

Participants received the development data at the beginning of May 2013, and the evaluation data at the beginning of June 2013. Results had to be submitted back to the organizers by September 9th 2013. Close to the deadline, some teams requested some more time, so we set an extended deadline on September 15th and marked the submissions arriving between both deadlines as late.

3. THE SWS 2013 MULTILINGUAL DATABASE

The database used for the SWS 2013 evaluation was collected thanks to a joint effort from several participating institutions that provided search utterances and queries on multiple languages and acoustic conditions (see Table 1). The database is available to the community for research purposes[1].

Table 1. Database contents disaggregated per language.

Language	data to search in (minutes / #utts)	#queries (dev / eval)	type of speech
Albanian	127 / 968	50 / 50	read
Basque	192 / 1.841	100 / 100	broadcast / read
Czech	252 / 3.667	94 / 93	conversational
Isixhosa	65 / 395	25 / 25	read
Isizulu	59 / 395	25 / 25	read
NNEnglish	141 / 434	61 / 60	lecture
Romanian	244 / 2.272	100 / 100	read
Sepedi	69 / 395	25 / 25	read
Setswana	51 / 395	25 / 25	read
Total	1.196 / 10.762	505 / 503	mixed

According to the spoken language and the recording conditions, the database is organized into 5 subsets:

African - 4 African languages: Isixhosa, Isizulu, Sepedi and Setswana. Recordings come from the Lwazi Corpus [7]. All 4 languages were recorded in similar acoustic conditions and contribute equally both to the search repository and the two sets of queries. All files include read speech recorded at 8 kHz through a telephone channel. Queries were obtained by cutting segments from speech utterances not included in the search repository. This subset features speaker mismatch but not channel mismatch between the search utterances and the queries.

Albanian & Romanian - Recordings come from the University Politehnica of Bucharest (SpeeD Research Laboratory). All files include read speech recorded through common PC microphones, originally at 16 kHz and then downsampled to 8 kHz to keep consistency with other subsets. Queries were obtained by cutting segments from speech utterances not included in the search repository. This subset features speaker mismatch and some channel mismatch between the search utterances and the queries, since different microphones on different PCs were used in recordings.

Basque - Speech utterances in the search repository come from the recently created Basque subset of the COST278 Broadcast News database [27], whereas the queries were specifically recorded for this evaluation. COST278 data include TV broadcast news speech (planned and spontaneous) in clean (studio) and noisy (outdoor) environments, originally sampled at 16 kHz and downsampled to 8 kHz for this evaluation. Three examples per query were read by different speakers and recorded in an office environment using a Roland Edirol R09

[1] We will provide the url on the camera ready version.

digital recorder. The Basque subset features both channel and speaker mismatch between the search utterances and the queries.

Czech - This subset contains conversational (spontaneous) speech obtained from telephone calls into radio live broadcasts, recorded at 8 kHz. The fact that all the recordings contain telephone-quality (i.e. low-quality) speech makes this subset more challenging than others in the database. Queries (10 examples per query, most of them from different speakers) were automatically cut (by forced alignment) from speech utterances not included in the search repository. This subset features speaker mismatch between the search utterances and the queries.

Non-native English - This subset includes lecture speech in English obtained from technical conferences in SuperLectures.com, speakers ranging from native to strong-accented non-native. Originally recorded at 44 kHz, audio files were downsampled to 8 kHz to keep consistency with other subsets. Queries were automatically extracted (by forced alignment) from speech utterances not included in the search repository. The original recordings were made using a high-quality microphone placed in front of the speaker, but might contain strong reverberation and some far-field channel effects. Therefore, besides speaker mismatch, there could be some channel mismatch between the search utterances and the queries.

The 9 languages selected for this database cover European and African language families. As a special case, the non-native English database consists of a mixture of native and non-native English speakers presenting their oral talks at different events. This subset thus presents a large variability in pronunciations, as it includes, for example, strong Indian English, French English and Chinese English, among others. Another interesting aspect of the database is the variety of speaking styles (read, planned, lecture, spontaneous) and the variety of acoustic (environment/channel) conditions, which forces systems to be built with low/zero resource constraints. The Basque subset is a good example of such mentioned variability, with read-speech queries recorded in an office environment and a set of search utterances extracted from TV broadcast news recordings including planned and spontaneous speech from a completely different set of speakers.

4. PERFORMANCE METRICS

In the SWS 2013 evaluation, four different performance metrics were used, measuring the detection accuracy and the computational resources required by the systems. As in previous SWS evaluations, the Actual Term Weighted Value (ATWV) was used as the primary metric, the other metrics being secondary or complementary. Note that ATWV is also the reference metric in NIST Spoken Term Detection evaluations [10] [20]. A new ATWV working point was defined, given by a prior that approximately matches the actual prior in the SWS 2013 search repository, and two suitable false alarm and miss error costs: $P_{\text{target}} = 0.00015$, $C_{\text{fa}} = 1$ and $C_{\text{miss}} = 100$. As usual, the Maximum Term Weighted Value (MTWV) —the highest value that can be attained by applying a single threshold to system scores— was also provided in order to evaluate miscalibration issues. Though not useful in a practical setting, the Upper Bound Term Weighted Value (UBTWV) —the highest value that can be attained if a different threshold per query is applied to system scores— was also computed in order to evaluate score normalization issues. Note that if the UBTWV score for a given system is much higher than the MTWV score, it means that scores are highly variable from query to query and thus a single threshold cannot optimize the performance *simultaneously* for all of them.

4.1. Normalized cross entropy metric

For the first time in a STD task, system performance was also evaluated in terms of the so called *normalized cross-entropy cost*, C_{nxe}, which is only based on system scores, in contrast to TWV, which evaluates system decisions. C_{nxe} measures the fraction of information, with regard to the ground truth, that is *not* provided by system scores, assuming that they can be interpreted as log-likelihood ratios. A perfect system would get $C_{\text{nxe}} \approx 0$ and a non-informative system would get $C_{\text{nxe}} = 1$, whereas $C_{\text{nxe}} > 1$ would indicate a severe miscalibration of the log-likelihood ratio scores (see [22] for details). The C_{nxe} was first introduced this year as an attempt to evaluate whether such a metric can be used as a feasible alternative to the ATWV metric, which has received many criticisms over the years, due to the embedded working point decisions it is build upon.

It must be noted that C_{nxe} is computed on system scores for a set of *trials*. Each trial consists of a query q and a segment x. For each trial, the ground truth is *True* or *False* depending on whether q actually appears in x or not. However, in a QbE-STD task, a system outputs scores only for a reduced subset of all the possible trials. But in order to compare the performance of two systems, they must refer to the *same* set of trials, usually the whole set of trials. Therefore, the evaluator must do a reasonable guess of the missing scores. It seems fair to assume that the missing scores are lower than the minimum submited score. In SWS 2013, all the missing trials by any given system were thus assigned the minimum score submitted by that system. However, this choice led to the unexpected result that C_{nxe} performance improved as the number of scores provided by a system increased, because as we consider additional trials most of them are false alarms and system scores are in most cases lower (that is, better) than the value that we would assign them if missing. For future SWS evaluations, this issue should be suitably addressed and a different *flavor* of C_{nxe} should be used, avoiding the above described bias.

Finally, since C_{nxe} measures both discrimination and calibration, a linear transformation minimizing C_{nxe} on the development set of queries was estimated in order to get $C_{\text{nxe}}^{\text{min}}$ and thus the calibration loss (again, see [22] for details).

4.2. Computational requirements

The computational requirements of the submitted systems, along with a description of the computing hardware (CPU model, RAM, OS, etc.), were self-reported by participants when returning their system results. As stated in [22], computational requirements were measured in terms of processing time and memory: the Real-Time (RT) factor and the Peak Memory Usage (PMU) were expected to be reported for both indexing (if needed) and searching. The RT factor involves two terms: (1) the Indexing Speed Factor (ISF), defined as the ratio of the indexing time to the source signal duation; and (2) the Searching Speed Factor (SSF), defined as the ratio of the total time employed in processing and searching the set of queries in the search repository to the product of their durations. In both cases, the total CPU time had to be reported as if all the computations were made in a single CPU. Two PMU figures were also defined in [22], corresponding to the indexing and searching phases, respectively. Most

teams, however, reported a single RT factor and a single PMU figure per system, usually corresponding to the searching phase.

5. SWS 2013 EVALUATION RESULTS

5.1. Overview of the submitted systems

In SWS 2013, 13 teams [1, 3, 5, 11, 8, 9, 12, 14, 15, 23, 25, 28, 29] submitted their system outputs for scoring. From these, 9 teams developed their primary system using frame-based approaches, which, in most cases, applied some flavor of the Dynamic Time Warping (DTW) algorithm [13], whereas 2 teams relied only on some form of symbol-based approach by using an Acoustic Key-Word Spotting (AKWS) algorithm [24]. Finally, 2 of the teams (BUT and L2F) combined frame-based and symbol-based algorithms, allowing them to achieve some of the best results in the evaluation. These systems provided either different ways of modeling the same information (e.g. BUT used the same features for DTW and AKWS subsystems) or different information sources under the same approach (e.g. BUT used 13 different phone decoders to extract features).

Although the aforementioned algorithms are all well known in the literature, every SWS evaluation brings forward some interesting ideas that combined with well-known techniques are able to achieve improvements. Aside from the fusion of multiple parallel sub-systems, the BUT system [25] also proposed a novel normalization technique (called M-norm) [26], in order to reduce the mismatch between scores from different queries. In the DTW implementation by L2F [1], a two-step approach was proposed that first performed a fast pass to find matching candidates, and then analyzed those candidates in more detail. A similar approach was followed by TID [5], with a first step based on the recently proposed IRDTW algorithm [6]. SpeeD [9] also proposed a DTW string matching algorithm, including a novel scoring normalization technique. Although the use of posterior probability features is well extended in the community, some variations included the use of articulatory bottleneck features by the IIIT-H team [15] and i-vectors by the LIA team [8]. It is also worth mentioning the tokenizer based on Gaussian component clustering that CUHK [29] implemented to get posterior probability vectors. Also from the CUHK team, we highlight the use of PSOLA to create 3 different-size queries prior to matching. Finally, it is interesting to note the introduction by the GT team [3] of a low-resource speech modeling algorithm using EHMM Models.

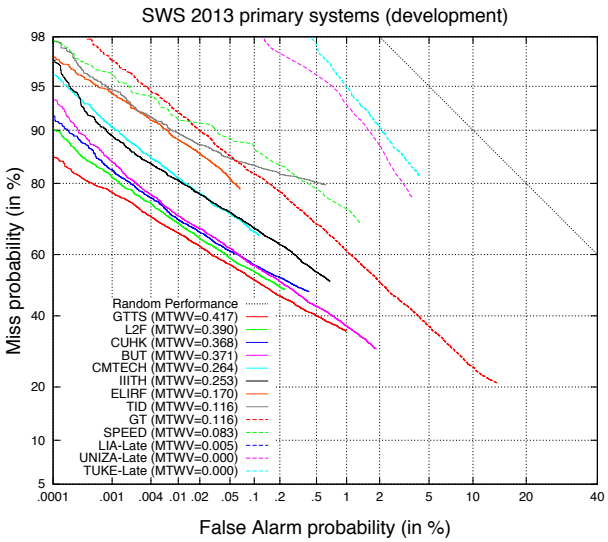

Fig. 1. DET curves for the primary systems on the development set.

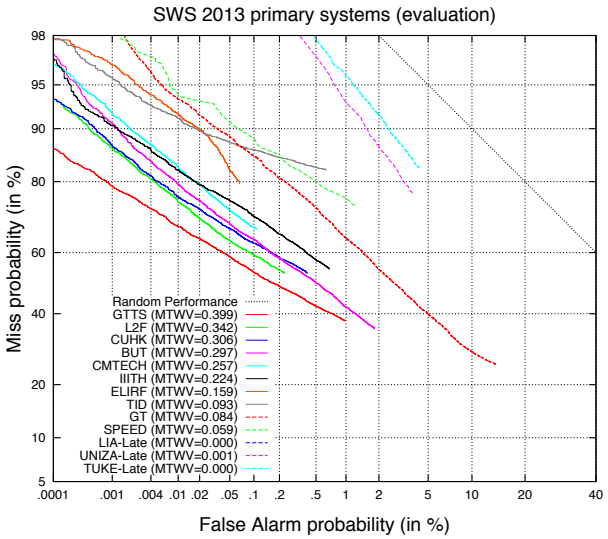

Fig. 2. DET curves for the primary systems on the evaluation set.

5.2. Analysis of performance

Figures 1 and 2 show the TWV DET curves for the primary systems submitted to SWS 2013 on the development and evaluation sets of queries, respectively. Each system is identified by a short team identifier or acronym, accompanied by the MTWV performance (for most systems, ATWV was close to MTWV). Please refer to the system papers listed in the references section to obtain more information on each system. The *Late* suffix indicates that the system was sent after the established deadline. The system labelled as *primary* was not necessarily the best performing system from a given team, though it usually was. We can see that none of the curves cover the full range of possible false alarm vs. miss probabilities, due to teams usually trimming the number of detections to lower their false alarm ratio, which is one of the big sources of error in the ATWV metric.

In some cases, the performance on the evaluation set did not degrade significantly with regard to the development set (e.g. for CMTECH and GTTS). However, in other cases (e.g. for BUT, CUHK and L2F) there was a remarkable degradation, revealing over-fitting issues which are difficult to explain. For instance, GTTS and L2F employed the same calibration and fusion approach and showed quite similar performance on the development set (on which calibration and fusion parameters were optimized), but L2F suffered a strong degradation on the evaluation set while GTTS did not.

Four of the five top-performing systems combined several sources of information: the GTTS system combined 4 DTW systems based on different phone posterior features; L2F combined an AKWS system and a DTW system; BUT combined 13 DTW and 13 AKWS systems, based on the same feature sets; and CMTECH performed an early combination of two kinds of features within the same DTW algorithm. Generally speaking, DTW-based algorithms (remarkably, GTTS) performed better than AKWS algorithms on the SWS 2013 datasets. The good performance of DTW systems could be partly due to the robustness of the set of features and the effectiveness of the fusion in extracting complementary information from several DTW-based subsystems (each based on a different set of features). Two of the best performing systems (L2F and BUT) used both DTW and AKWS algorithms. In both cases, DTW systems got

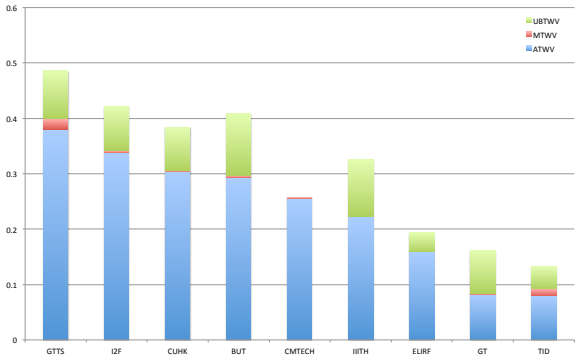

Fig. 3. ATWV, MTWV and UBTWV results on eval queries for primary systems with positive ATWV scores

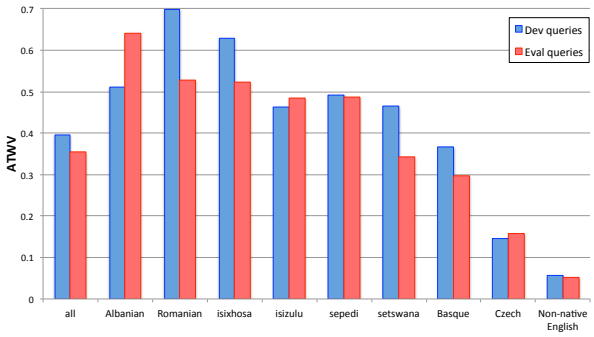

Fig. 4. Average ATWV per language (10 best performing systems).

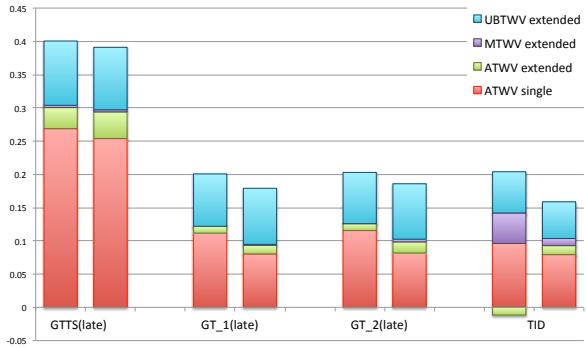

Fig. 5. TWV performance for systems using multiple examples per query (left bar: development, right bar: evaluation).

better performance than AKWS systems. Moreover, the BUT team used the same sets of features and the same score normalization and fusion approaches for both DTW and AKWS systems. On the other hand, BUT reported that AKWS performed better than DTW on subsets with stronger acoustic mismatch (Basque and non-native English). Based on these results, we may say that DTW performs slightly better than AKWS, but the best choice would probably be combining both types of systems.

Figure 3 shows the stacked ATWV, MTWV and UBTWV scores for systems with positive ATWV scores. We see how, in general, ATWV scores are very close to MTWV scores, which means that systems are able to generalize well once their parameters are tuned on the development set. In addition, we see how CMTECH obtained a UBTWV score almost equal to the MTWV. This is due to the low number of results that this team returned (i.e. only those that were clear matches).

Figure 4 shows the average ATWV for the 10 best-performing systems overall (i.e. including both primary and contrastive, either on-time or late submissions) on the 9 language-specific subsets contained in the database. As may be expected, best performance was obtained on subsets containing high-quality recordings in a lab environment (Albanian and Romanian), while the worst was obtained, by far, on non-native English, which featured reverberant and relatively far-distance recordings with highly variable pronunciations. Results for South-African languages were on the average (slightly better for Isixhosa and slightly worse for Setswana). In the case of Basque, systems attained lower performance than expected, probably due to a strong mismatch between the search utterances and the queries. Results for Czech were even worse, which was quite surprising, since the search utterances and the queries featured the same acoustic conditions, which were not tremendously challenging. A possible explanation could be that Czech conversational speech can be really fast, which caused queries to be quite short when cut from actual conversations by using forced alignment, with no silence around them. In fact, a Czech native speaker was able to recognize those short queries only after listening to the whole sentences where they appeared.

Figure 5 shows the TWV performance for systems that processed multiple examples per query (when available). This subtask was a novelty in SWS 2013 and only 3 teams submitted systems for it (GTTS, GT and TID). Only two languages provided multiple examples per query: Basque and Czech, with up to 3 and 10 examples per query, respectively. However, results in Figure 5 are shown as evaluated for the whole database, since no information on what language each query came from was given to participants. ATWV scores are shown for systems using a single example (ATWV_single) and systems using all the available examples per query (ATWV_extended). Besides, the Maximum TWV and the Upper Bound TWV scores are also shown in the extended condition. In most cases (remarkably, for GTTS), using multiple examples per query did improve performance. The only exception was the TID system on the development set of queries.

Figure 6 shows the performance of primary systems in terms of the newly proposed C_{nxe} metric. Both the actual and the minimum C_{nxe} values are shown. It must be noted that some systems obtained a very bad actual C_{nxe}, in part due to bad calibration, but also to the issue mentioned in Section 4. Some systems returned a small number of detections in order to minimize the risk of increasing the number of false alarms (which greatly penalizes the ATWV metric). But the computation of C_{nxe} requires a score for each possible trial, so that missing trials (those for which the system does not output a score) are assigned a default score. When the number of system detections is very small, this can result in a non-informative C_{nxe} value. On the other hand, systems returning a relatively high number of detections (e.g. GT) attained a good result in terms of C_{nxe}, since the scores provided by the system were better (on average) than those assigned by default. As noted in Section 4, though this metric may eventually replace the TWV metrics in future evaluations, the issue of missing trials must be suitably addressed (or the task re-defined, so that systems provide scores for all the trials) for comparisons among systems to be fair.

Finally, Table 2 shows the computational requirements —real-time (RT) factor and peak memory usage (PMU)— of the primary systems submitted to SWS 2103. These values were self-reported by

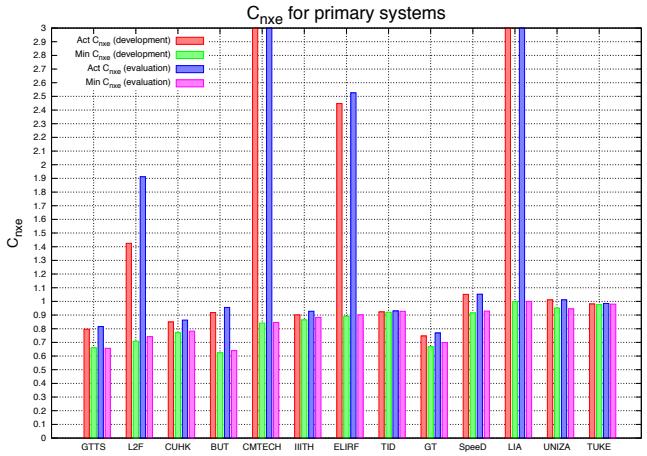

Fig. 6. Actual and Minimum C_{nxe} performance for primary systems on the development and evaluation sets of queries.

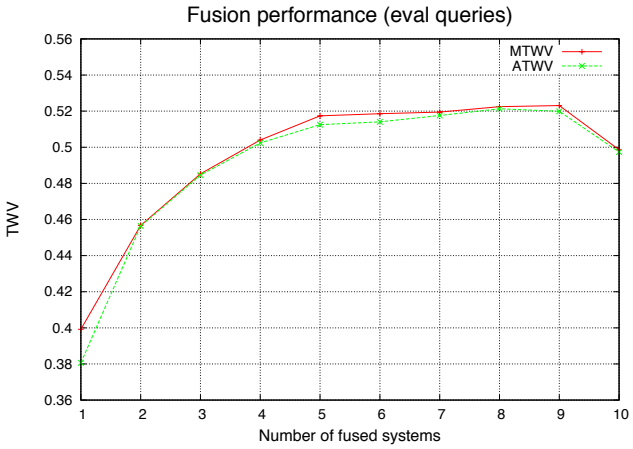

Fig. 7. Fusion performance (10 best primary systems).

Table 2. Self-reported Real-Time (RT) factor and Peak memory Usage (PMU) for the primary systems submitted to SWS 2013.

Team name	RT	PMU (MB)	Approach
UNIZA [14]	2.2E-2	10	Symbol
GT [3]	3.0E-3	50	Symbol
L2F [1]	3.4E-1	75	Frame+Symbol
TID [5]	1.3E-4	110	Frame
GTTS [23]	1.7E-2	200	Frame
BUT [25]	2.0E-1	210	Frame+Symbol
LIA [8]	1.3E-4	1229	Frame
TUKE [28]	4.8E-3	1843	Frame
SPEED [9]	6.0E-5	7270	Frame
IIIT-H [15]	4.1E-4	10240	Frame
CUHK [29]	1.8E-2	10240	Frame
CMTECH [12]	5.6E-3	11776	Frame
ELIRF [11]	2.8E-3	12288	Frame
Mean	8.4E-2	2700	–

participants, suing different machines and procedures, so strong conclusions cannot be drawn from them. Generally speaking, the PMU for pure symbol-based systems was much smaller than that of frame-based systems, simply because the former just need to load the necessary models in memory to conduct Viterbi (or similar) decoding, instead of storing similarity matrices and performing dynamic programming. Among systems using DTW-based algorithms, GTTS, BUT and TID reported competitive memory requirements. In particular, TID DTW-like implementation [5] was designed to avoid storage of any similarity matrix. On the other hand, RT values are usually smaller for frame-based systems. An exception to this is the GT system, which uses an Ergodic-HMM model which is able to generate a 3D lattice structure, whose speed is above average for symbol-based systems. In general, we believe that RT values must be greatly improved to make QbE-STD search on real-life data interesting for commercial applications.

6. FUSION STUDY

Inspired by the improvements in performance attained by some of the participants when fusing systems based on different algorithms or features, a late (score-level) fusion study was performed by incrementally fusing the 10 best-performing primary systems, under the calibration/fusion approach described in [2], which was successfully applied by GTTS and L2F in their submissions [23] [1].

The fusion procedure first aligns the detections of several systems, then retains some of them through majority voting and finally hypothesizes the scores for any missing trials (typically by using the minimum system score per query). In this way, the original STD task is converted into a verification task. Then, like in other verification tasks, a linear combination of system scores is estimated on the development set through linear logistic regression. As a result, the combined scores are well calibrated and the optimal Bayes detection threshold, given by the application parameters (prior and costs), is applied (see [2] for details).

Figure 7 shows the ATWV/MTWV evolution on the evaluation set when fusing the N best primary systems, for $N = 2, 3, \ldots, 10$. Systems were fused in order of performance (see Fig. 2). The performance for the best individual system is shown too ($N = 1$). Most of the improvement was already obtained for $N = 5$, but ATWV kept improving until $N = 8$ (ATWV: 0.5213) and the best MTWV was obtained for $N = 9$ (MTWV: 0.5231), meaning a 30% relative improvement over the best individual system. A more in-depth study of fusions is planned which will try all the combinations of systems or a greedy selection approach such as that proposed in [21], in order to determine which kind of systems are worth fusing.

7. CONCLUSIONS AND FUTURE PERSPECTIVES

The Spoken Web Search (SWS) task, carried out within the MediaEval benchmark campaign, consists of finding instances of a spoken query in a set of spoken documents (the search repository). The speech database supporting the evaluation typically features several low-resource languages and includes recordings under different (sometimes challenging) acoustic conditions. Participants must build systems that can cope with this variability without knowing what language each utterance corresponds to. This means that systems must be designed for a low-resource setting.

For the SWS 2013 evaluation (the third in the series), a database was prepared consisting of a search repository of around 20 hours, with more than 10.000 utterances, and two sets of more than 500 spoken queries. Speech data were recorded through different types of channels in different environments and featured 9 different languages. A record in participation was attained, with 13 teams sub-

mitting at least one system.

In this paper, besides presenting the setup, datasets and performance measures of the SWS 2013 evaluation, we have analyzed the results obtained by the submitted systems and presented a post-evaluation study where the 10 best-performing systems were incrementally fused (at the score level), obtaining a 30% relative improvement over the best-performing individual system, proving the benefits of combining independent or complementary sources of information or different modeling approaches.

Given the increasing interest for this task in the community, we are already planning a new edition of the SWS evaluation, renamed QUESST, i.e. *Query by Example Spoken Search Task*, within the Mediaeval 2014 benchmark campaign. This year, we will continue tackling the problem of low-resource settings and will introduce a component of variability between queries and references, allowing for a limited amount of acoustic insertions to still be considered matches.

8. ACKNOWLEDGEMENTS

We would like to thank Charl Van Heerden for his help in preparing the datasets for African languages. We would also like to thank Martha Larson and Gareth Jones for organizing the Mediaeval benchmark evaluation.

9. REFERENCES

[1] Alberto Abad, Ramon F. Astudillo and Isabel Trancoso, "The L2F Spoken Web Search system for Mediaeval 2013", in *Proc. Mediaeval 2013 Workshop*, Barcelona, Spain, 2013.

[2] Alberto Abad, Luis J. Rodriguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, G. Bordel, "On the Calibration and Fusion of Heterogeneous Spoken Term Detection Systems", in *Proc. Interspeech 2013*, Lyon, France, August 25-29, 2013.

[3] Asif Ali and Mark A Clements, "Spoken Web Search using an Ergodic Hidden Markov Model of Speech", in *Proc. Mediaeval 2013 Workshop*, Barcelona, Spain, 2013.

[4] Xavier Anguera, Florian Metze, Andi Buzo, Igor Szoke and Luis Javier Rodriguez-Fuentes, "The Spoken Web Search Task", in *Proc. Mediaeval 2013 Workshop*, Barcelona, Spain, 2013.

[5] Xavier Anguera, Miroslav Skácel, Volker Vorwerk and Jordi Luque, "The Telefonica Research Spoken Web Search System for MediaEval 2013", in *Proc. Mediaeval 2013 Workshop*, Barcelona, Spain, 2013.

[6] Xavier Anguera, "Information Retrieval-based Dynamic Time Warping", in *Proc. Interspeech*, Lyon, France, 2013.

[7] E. Barnard, M. Davel and C. van Heerden, "ASR corpus design for resource-scarce languages", in *Proc. Interspeech 2009, pp. 2847–2850*, Brighton, UK, September 2009.

[8] Mohamed Bouallegue, Grégory Senay, Mohamed Morchid, Driss Matrouf and Richard Dufour, "LIA @ MediaEval 2013 Spoken Web Search Task : An I-Vector based Approach", in *Proc. Mediaeval 2013 Workshop*, Barcelona, Spain, 2013.

[9] Andi Buzo, Horia Cucu, Iris Molnar, Bogdan Ionescu and Corneliu Burileanu, "SpeeD @ MediaEval 2013 : A Phone Recognition Approach to Spoken Term Detection", in *Proc. Mediaeval 2013 Workshop*, Barcelona, Spain, 2013.

[10] Jonathan G. Fiscus, Jerome Ajot, John S. Garofolo and George Doddington, "Results of the 2006 Spoken Term Detection Evaluation", in *Proc. SIGIR 2007 Workshop on Searching Spontaneous Conversational Speech, pp. 51–57*, Amsterdam, 2007.

[11] Jon A. Gómez, Lluís-F. Hurtado, Marcos Calvo and Emilio Sanchis, "ELiRF at MediaEval 2013 : Spoken Web Search Task", in *Proc. Mediaeval 2013 Workshop*, Barcelona, Spain, 2013.

[12] Ciro Gracia, Xavier Anguera and Xavier Binefa, "The CMTECH Spoken Web Search System for MediaEval 2013", in *Proc. Mediaeval 2013 Workshop*, Barcelona, Spain, 2013.

[13] Timothy J. Hazen, Wade Shen adn Christopher White, "Query-By-Example Spoken Term Detection Using Phonetic Posteriorgram Templates", in *Proc. IEEE ASRU Workshop 2009, pp. 421–426*.

[14] Roman Jarina, Michal Kuba, Róbert Gubka, Michal Chmulik and Martin Paralic, "UNIZA System for the Spoken Web Search Task at", in *Proc. Mediaeval 2013 Workshop*, Barcelona, Spain, 2013.

[15] Gautam Mantena and Kishore Prahallad, "IIIT-H SWS 2013 : Gaussian Posteriorgrams of Bottle-Neck Features for Query-by-Example Spoken Term Detection", in *Proc. Mediaeval 2013 Workshop*, Barcelona, Spain, 2013.

[16] Florian Metze, Nitendra Rajput, Xavier Anguera, Marelie Davel, Guillaume Gravier, Charl van Heerden, Gautam V. Mantena, Armando Muscariello, Kishore Prahallad, Igor Szoke and Javier Tejedor, "The Spoken Web Search Task at MediaEval 2011", in *Proc. ICASSP 2012, pp. 5165–5168*, Kyoto, Japan, March 25-30, 2012.

[17] Florian Metze, Eric Fosler-Lussier and Rebecca Bates, "The Speech Recognition Virtual Kitchen", in *Proc. Interspeech 2013, pp. 1858–1860*, Lyon, France, August 25-29, 2013.

[18] Florian Metze, Xavier Anguera, Etienne Barnard, Marelie Davel and Guillaume Gravier, "The Spoken Web Search Task at MediaEval 2012", in *Proc. ICASSP 2013*, pp. 8121–8125, Vancouver, Canada, May 26-31, 2013.

[19] Florian Metze and Xavier Anguera and Etienne Barnard and Marelie Davel and Guillaume Gravier, "Language Independent Search in MediaEval's Spoken Web Search Task", in IEEE Journal on Computer Speech and Language, Special Issue on Information Extraction & Retrieval. To appear.

[20] NIST Open Keyword Search 2013 Evaluation (OpenKWS13), "OpenKWS13 Keyword Search Evaluation Plan", March 8, 2013. http://www.nist.gov/itl/iad/mig/upload/OpenKWS13-EvalPlan.pdf

[21] Luis J. Rodriguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, G. Bordel, D. Martinez, J. Villalba, A. Miguel, A. Ortega, E. Lleida, A. Abad, O. Koller, I. Trancoso, P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, R. Saeidi, M. Soufifar, T. Kinnunen, T. Svendsen, P. Franti, "Multi-site Heterogeneous System Fusions for The Albayzin 2010 Language Recognition Evaluation", in *Proc. IEEE ASRU Workshop*, Hawaii, USA, December, 2011.

[22] Luis J. Rodriguez-Fuentes and Mikel Penagarikano, "MediaEval 2013 Spoken Web Search Task: System Performance Measures", *Technical Report-2013-1, Dept. Electricity and Electronics, University of the Basque Country*, May 30, 2013, http://gtts.ehu.es/gtts/NT/fulltext/rodriguezmediaeval13.pdf

[23] Luis J. Rodriguez-Fuentes, Amparo Varona, Mikel Penagarikano, Germán Bordel and Mireia Diez, "GTTS Systems for the SWS Task at MediaEval 2013", in *Proc. Mediaeval 2013 Workshop*, Barcelona, Spain, 2013.

[24] Igor Szöke, Petr Schwarz, Pavěl Matejka, Lukáš Burget, Martin Karafiát, Michal Fapšo and Jan Černocký, "Comparison of Keyword Spotting Approaches for Informal Continuous Speech", in *Proc. Interspeech, pp. 633–636*, Lisbon, Portugal, 2005.

[25] Igor Szöke, Lukáš Burget, František Grézl and Lucas Ondel, "BUT SWS 2013 - Massive Parallel Approach", in *Proc. Mediaeval 2013 Workshop*, Barcelona, Spain, 2013.

[26] Igor Szöke, Lukáš Burget, František Grézl, Jan "Honza" Černocký and Lucas Ondel, "Calibration and Fusion of Query-by-Example systems - BUT SWS 2013", in *Proc. IEEE ICASSP*, Florence, Italy, May 4-9, 2014.

[27] An Vandecatseye, Jean-Pierre Martens, Joao Neto, Hugo Meinedo, Carmen Garcia-Mateo, Javier Dieguez, France Mihelic, Janez Zibert, Jan Nouza, Petr David, Matus Pleva, Anton Cizmar, Harris Papageorgiou and Christina Alexandris, "The COST278 pan-European Broadcast News Database", in *Proc. LREC 2004, pp. 873-876*, Lisbon, 2004.

[28] Jozef Vavrek, Matúš Pleva, Martin Lojka, Peter Viszlay, Eva Kiktová, Daniel Hládek, Jozef Juhár, Matus Pleva, Eva Kiktova, Daniel Hladek, and Jozef Juhar, "TUKE at MediaEval 2013 Spoken Web Search Task", in *Proc. Mediaeval 2013 Workshop*, Barcelona, Spain, 2013.

[29] Haipeng Wang and Tan Lee, "The CUHK Spoken Web Search System for MediaEval 2013", in *Proc. Mediaeval 2013 Workshop*, Barcelona, Spain, 2013.