Query-by-Example Spoken Term Detection on Multilingual Unconstrained Speech

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 audio queries in a low-resource setting. After two successful editions and a continuously growing interest in the scientific community, a special effort was made in SWS 2013 to prepare a challenging database, including speech in 9 different languages with diverse environment and channel conditions. In this paper, first we describe the database and the performance metrics. Then, we briefly review the algorithmic approaches followed by participants and present and discuss the obtained performances, which demonstrate the feasibility of the proposed task, even under such challenging conditions (multiple languages and unconstrained acoustic conditions). Finally, we analyze the fusion of the top-performing systems, which achieved a 30% relative improvement over the best single system in the evaluation, proving 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, pattern matching

1. Introduction

The MediaEval benchmark evaluation proposes every year a set of tasks on multimedia analysis. The Spoken Web Search (SWS) 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 [9] and, more recently, the OpenKWS13 evaluation [19], 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 zero-resource approaches have to be employed.

*Igor Szöke was supported by Grant Agency of Czech Republic post-doctoral project No.GPP202/12/P567.

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. In this regard, to make life easier to newcomers, starting from SWS 2012 a handful of systems and features has been released through the so called *Speech Recognition Virtual Kitchen* [16].

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 [15], containing around 3 hours of spontaneous telephone voice messages in 4 languages spoken in India. A different dataset was used for the SWS 2012 evaluation, consisting of around 8 hours of speech recordings (4 hours for development and 4 hours for evaluation) of 4 languages spoken in Africa [17]. See [18] for a comprehensive analysis of techniques proposed in these 2 years. For the SWS 2013 evaluation, a single set of speech utterances was provided to perform the search of queries, with around 20 hours of speech in 9 different languages, which is more than twice the size of the search datasets used in SWS 2012. The number of queries also increased remarkably, from 100 queries in SWS 2012 to more than 500 queries in SWS 2013.

2. The SWS 2013 Multilingual Database

The database used for the SWS 2013 evaluation has been 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].

Unlike in previous SWS evaluations, a unique set of speech utterances was used here to search for queries. Two sets of queries were defined, one for tuning (development) the systems and the other for measuring system performance (evaluation). The mixture of languages and acoustic conditions in the search repository was so large that trying to adapt a system to those

[1]http://speech.fit.vutbr.cz/files/sws2013Database.tgz

Table 1: Database contents disaggregated per language.

Language	data to search in (min / #seg)	#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
NNEnglish	141 / 434	61 / 60	lecture
Romanian	244 / 2.272	100 / 100	read
Isixhosa	65 / 395	25 / 25	read
Isizulu	59 / 395	25 / 25	read
Sepedi	69 / 395	25 / 25	read
Setswana	51 / 395	25 / 25	read
Total	1.196 / 10.762	505 / 503	mixed

conditions was not only acceptable but an interesting issue to do research on. Note that, given that utterances in the search repository were shuffled and no side information was provided to participants regarding the spoken language or the acoustic conditions, any possible form of adaptation would have to rely on unsupervised algorithms, thereby introducing an interesting line of research.

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 [6]. All 4 languages were recorded in similar acoustic conditions and contribute equally both to the search repository and the 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 [25], 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 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 strongly accepted english (such as e.g. Indian, French and Chinese accented 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, 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.

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). Participants did not know whether those queries all came from the same or from different languages. In addition to a basic (required) submission involving a single example per query to perform the search, participants were invited to carry out an extended series of runs where they could use all the available examples per query.

3. 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. Note that ATWV is also the reference metric in NIST Spoken Term Detection evaluations [9] [19]. 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 $C_{\text{miss}} = 100$ (see [21] for details). 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 misscalibration 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 UBTWV is much better than MTWV, 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.

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 $C_{\mathrm{nxe}} \approx 0$ and a non-informative system would get $C_{\mathrm{nxe}} = 1$, whereas $C_{\mathrm{nxe}} > 1$ would indicate a severe miscalibration of the log-likelihood ratio scores (see [21] for details).

Computational requirements were measured in terms of the real-time factor and the peak memory usage for both indexing (if needed) and searching, and an overall processing load measure was also defined [21]. In this paper, performance analysis will focus on TWV metrics (ATWV, MTWV and DET curves).

4. SWS 2013 Evaluation Results

4.1. Overview of the submitted systems

In SWS 2013, 13 teams [1, 3, 5, 10, 7, 8, 11, 13, 14, 22, 23, 26, 27] submitted their system outputs for scoring. From these, 9 teams developed their primary system based on a Dynamic Time Warping (DTW) approach [12], while only 2 teams relied only on some form of Acoustic Key-Word Spotting (AKWS) [24]. Finally, 2 of the teams (BUT and L2F) combined DTW and AKWS algorithms, which allowed them to achieve some of the best results in the evaluation. Most of the best performing systems were in fact the fusion of several subsystems. These subsystems 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). Besides having a robust set of features (phone posteriorgrams were among the most common representations) and combining the detections of several subsystems, two key issues for achieving good performance were voice activity detection and score normalization.

4.2. Performance of individual (primary) systems

Due to a lack of space, this paper focuses on the primary systems submitted to the required condition (using a single example per query). Figures 1 and 2 show the TWV DET curves for the primary systems submitted to SWS 2013, on the sets of development and evaluation 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). 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.

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.

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

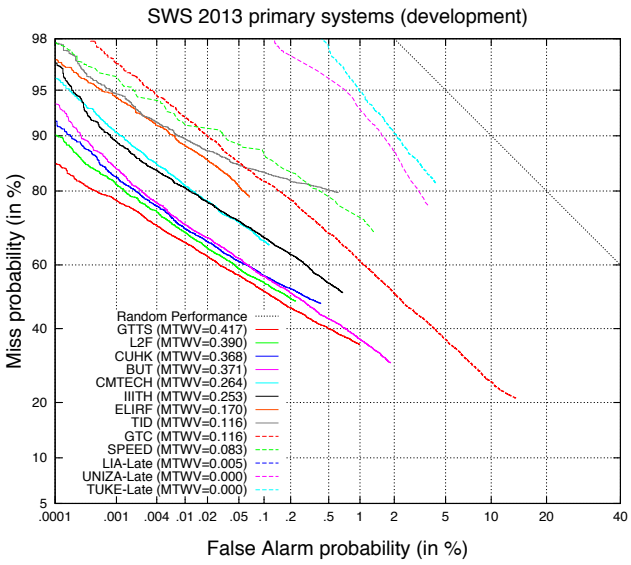

Figure 1: DET curves for the primary systems on the development set. Systems in the legend are ordered according to the MTWV.

DTW and 13 AKWS systems, based on the same feature sets; and CMTECH performed an early combination of two kinds of features within the 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 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 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. A possible explanation could be that Czech conversational speech can be really fast and the queries be cut out using forced alignment with no silence around them. The small length of some queries and the mismatch between fast and slow pronunciations might have also played a role. In fact, a Czech native speaker was able to

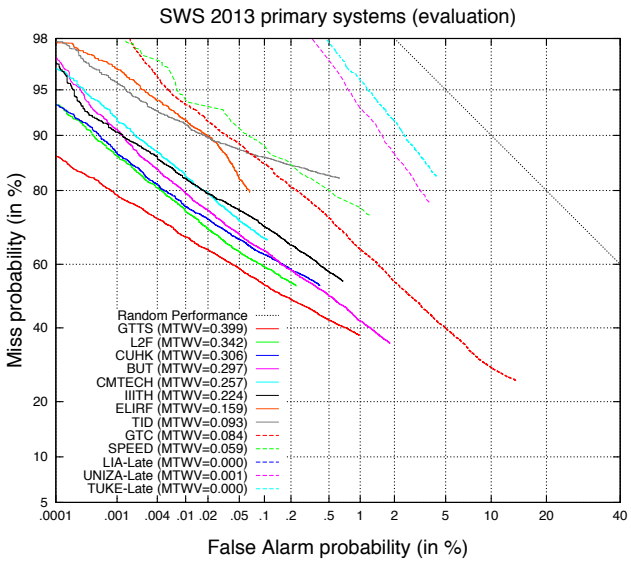

Figure 2: DET curves for the primary systems on the evaluation set. Systems in the legend are ordered according to the MTWV.

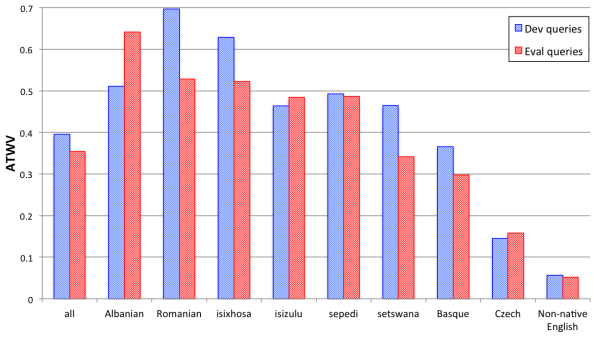

Figure 3: Average ATWV per language (10 best performing systems).

recognize those short queries only after listening to the whole sentences where they appeared.

Processing load and peak memory were self-reported by participants and so, many times, were not computed in equal conditions or using the same information in all systems. Peak memory usage for pure AKWS systems is much smaller than that of DTW-based systems, simply because the former just need to load the necessary models 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. Real-time factor averaged 0.05, ranging from $5e^{-5}$ and 0.2. In our opinion, these values should be greatly improved to make QbE-STD search on real-life data interesting for commercial applications.

4.3. Fusion performance

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 [22] [1].

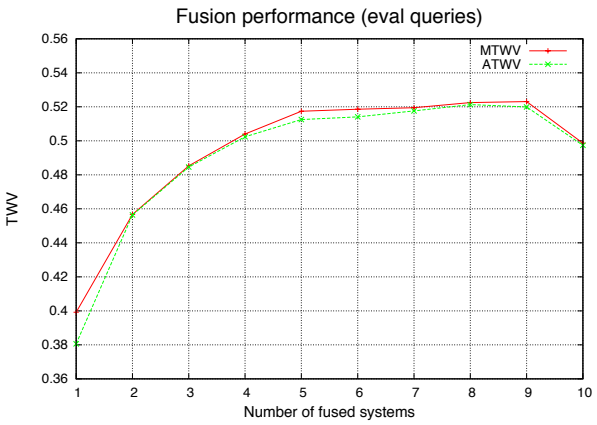

Figure 4: Fusion performance (10 best primary systems).

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 4 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 [20], in order to determine which kind of systems are worth fusing.

5. Conclusions

In this paper, the main features of the SWS evaluation at MediaEval 2013 have been presented and the obtained results have been briefly analyzed. A challenging database was created specifically for this evaluation, consisting of a search repository of around 20 hours of speech in 9 different languages, recorded in diverse acoustic conditions, and two sets of more than 500 spoken queries. A record in participation was attained, with 13 teams submitting systems. Results show that, even though the database proved quite challenging, most of the submitted systems could tackle the task and obtained very reasonable performances. A post-evaluation study of the incremental fusion of the 10 best-performing systems was carried out, obtaining a 30% relative improvement over the best-performing individual system, proving the benefits of combining heterogeneous sources of information or different modeling approaches.

6. 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.

7. 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] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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*.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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

[20] 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.

[21] 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

[22] 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.

[23] Igor Szöke, Lukáš Burget, František Grézl and Lucas Ondel, "BUT SWS 2013 - Massive Parallel Approach", 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] 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.

[26] 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.

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