QUESST2014: EVALUATING QUERY-BY-EXAMPLE SPEECH SEARCH IN A ZERO-RESOURCE SETTING WITH REAL-LIFE QUERIES

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

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

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

ABSTRACT

In this paper, we present the task and describe the main findings of the 2014 "Query-by-Example Speech Search Task" (QUESST) evaluation. The purpose of QUESST was to perform language independent search of spoken queries on spoken documents, while targeting languages or acoustic conditions for which very few speech resources are available. This evaluation investigated for the first time the performance of query-by-example search against morphological and morpho-syntactic variability, requiring participants to match variants of a spoken query in several languages of different morphological complexity. Another novelty is the use of the normalized cross entropy cost (C_{nxe}) as the primary performance metric, keeping Term-Weighted Value (TWV) as a secondary metric for comparison with previous evaluations. After analyzing the most competitive submissions (by five teams), we find that, although low-level "pattern matching" approaches provide the best performance for "exact" matches, "symbolic" approaches working on higher-level representations seem to perform better in more complex settings, such as matching morphological variants. Finally, optimizing the output scores for C_{nxe} seems to generate systems that are more robust to differences in the operating point and that also perform well in terms of TWV, whereas the opposite might not be always true.

Index Terms— low-resource speech recognition, query-by-example speech search, spoken term detection

1. INTRODUCTION

In today's globalized society, we have the opportunity to analyze acoustic material from minority cultures and languages, which are neither politically nor economically interesting for large corporations to develop speech technology for. The low- or zero-resource language processing area of research focuses on devising techniques to extract information from audio for which little (or no) information is available. Within the MediaEval benchmarking campaign [1], this is the fourth year that a task was organized to search for spoken queries within low-resourced spoken documents [2, 3, 4, 5, 6]. The "Query-by-Example Speech Search Task" (QUESST, formerly "SWS" or "Spoken Web Search" task) involved more than 1000 spoken queries, which had to be searched in a speech database of over 23 hours. Participants did not know which language a specific query or test utterance belonged to. Differently from previous years, given a spoken query q and a spoken document x, a detection score plus a discrete Yes/No decision were mandatorily required, but not the exact location(s) of q inside x. Submissions were received from 9 teams [7, 8, 9, 10, 11, 12, 13, 14, 15], who approached the search problem from multiple angles.

Similar efforts in this direction in the area include the Spoken Term Detection (STD) evaluation conducted by NIST in 2006 [16] and, more recently, the OpenKWS evaluations [17]. Unlike QUESST, these evaluations perform *exact* search on speech using *textual* input on a *single*, known, language for which well-transcribed training data is available. These evaluations therefore draw systems that are built specifically for a target language, and that are usually driven by traditional ASR techniques, even though they are often trained with (relatively) few available labeled training data.

In this paper, we describe and analyze the effect of two major changes introduced in this year's evaluation, which triggered also the change in name. On the one hand, the inclusion of three different kinds of query matches accounts for the expectations of a typical user of a voice search application. These are: finding exact occurrences of the spoken query, finding approximate occurrences (morphological variations) and finding parts of the spoken query, possibly in different order and with filler content inbetween (morpho-syntactic variations). On the other hand, the *normalized cross entropy cost* (C_{nxe}) replaces the Term-Weighted Value (TWV) as the main performance metric this year. As shown in Section 3, if systems are designed to minimize C_{nxe} (not to maximize TWV), target and non-target detection scores get more separated one from another, making systems work efficently on a wider range of operating points. Results reveal that systems optimized for C_{nxe} also perform well in terms of TWV, while the opposite is not always true.

At the conclusion of the evaluation, in order to foster research in this area within the research community, we made the search database, the queries and the ground truth files freely available for research purposes [18].

2. THE QUESST 2014 MULTILINGUAL DATABASE

The QUESST 2014 search dataset consists of 23 hours or around 12.500 spoken documents in the following languages: Albanian, Basque, Czech, non-native English, Romanian and Slovak [6]. The languages were chosen so that relatively little annotated data can be found for them, as would be the case for a "low resource" language.

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

The recordings were PCM encoded with 8 KHz sampling rate and 16 bit resolution (down-sampling or re-encoding were done when necessary to homogenize the database). The spoken documents (6.6 seconds long on average) were extracted from longer recordings of different types: read, broadcast, lecture and conversational speech. Besides language and speech type variability, the search dataset also features acoustic environment and channel variability. The distribution of spoken documents per language is shown in Table 1.

Table 1. Speech data distribution in QUESST 2014: spoken documents and queries, broken down for query subsets T1-T3, see below.

	Search dataset		Number (and subset) of queries	
	#docs / size (min)		dev (T1/T2/T3)	eval (T1/T2/T3)
Albanian	968 /	127	50 (20/13/16)	50 (18/13/17)
Basque	1841 /	192	70 (16/33/21)	70 (30/19/21)
Czech	2652 /	237	100 (77/24/27)	100 (73/27/32)
NNEnglish	2438 /	273	138 (46/46/46)	138 (46/46/46)
Romanian	2272 /	244	100 (46/21/31)	100 (43/27/30)
Slovak	2320 /	312	102 (102/53/14)	97 (97/47/10)
Total	12491 /	1385	560 (307/190/155)	555 (307/179/156)

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

Albanian & Romanian: Read speech by 10 Albanian and 20 Romanian speakers (gender balanced) in a lab environment. Queries were recorded by different speakers within the same environment.

Basque: Mixture of read and spontaneous speech recorded from broadcast news programs, including some studio and outdoors (noisy) recordings. Queries were recorded with a digital recorder (using a close-talk microphone) in an office environment by different speakers (gender balanced).

Czech: Conversational (spontaneous) speech obtained from telephone calls into radio live broadcasts, mixing some clean (majority) and noisy acoustic conditions. Queries were recorded by 12 speakers (3 non-native) using a mobile application (total acoustic mismatch).

Non-native English: The main corpus was compiled from a variety of TED talks [19] with non-native, but skilled English speakers. Transcriptions were automatically aligned with the audio to generate the references. Queries were spoken in clean conditions by non-native (Chinese, Indian, German, and Italian) English speakers of intermediate proficiency.

Slovak: Spontaneous speech, recorded from Parliament meetings using stationary microphones, in mainly clean conditions (90% male speakers). Queries were recorded in clean lab conditions.

In addition to the search data, two sets of audio queries were prepared: 560 queries for development, and 555 queries for evaluation. This year, the queries were not extracted (i.e. "cut") from longer recordings, in order to avoid imposing acoustic context. Instead, they were recorded manually and the recruited speakers were asked to pronounce the queries in isolation, at a normal speaking rate and using a clear speaking style to simulate a regular user querying a retrieval system via speech. Three types of matches were considered, those that most users could expect in a real-life scenario, in increasing order of complexity:

Type 1: Occurrences in the search data should match *exactly* the lexical form of the query. For example, the utterance "My white horse is beautiful" would match the query "white horse". This type of match is the same as in SWS2013 [5].

Type 2: Occurrences in the search data may contain small morphological variations with regard to the lexical form of the query. For example, inflectional forms, added or omitted prefixes and suffixes should be correctly matched with the given query. In all cases, the matching part of any query is set to be at least 5 phonemes (approximately 250 ms) long, whereas the non-matching part should be much smaller. For example, the utterance "There were too many researchers at the conference" would match the query "research".

Type 3: Occurrences in the search data may contain both syntactic and morphological differences (i.e. word reorderings + word inflections), along with some filler content, with regard to the lexical form of the query. For example, the utterances "In Stockholm the snow is whiter than here" and "Blue or white, the snow is cold" would both match the query "white snow". Note that there should not be any silence between words, as queries and utterances are pronounced fluently. Though this type of match may not be appropriate in all languages, the present work investigates the potential of search techniques to support this case, if required.

Three disjoint sets of queries were defined: T1, T2 and T3, according to the most complex type of match that was found in the search data. In the case of T1 queries, only Type 1 matches are found. In the case of T2 queries, Type 2 along with possibly some Type 1 matches are found. Finally, in the case of T3 queries, Type 3 along with possibly some Type 1 and Type 2 matches are found. The information about query subsets T1, T2 and T3 and the language spoken in audio files was not made available to participants during the evaluation, so that systems were implicitly required to be language-independent and to detect all possible types of matches.

3. EVALUATION OF RESULTS

3.1. Evaluation Metrics

The normalized cross entropy cost (C_{nxe}) was the primary metric in QUESST 2014, but TWV was also used in order to allow the comparison to previous evaluations, where TWV was the primary metric.

TWV is a well known metric defined by NIST [20] and used in the community to compare the accuracy of keyword spotting/spoken term detection systems. Actual TWV (ATWV) is calculated according to a per-query Yes/No decision assigned to every system detection to tell the scoring system whether the detection is believed to be a hit or a false alarm (FA). Maximum TWV (MTWV) can be calculated by searching for the global threshold (to set Yes/No decisions) that maximizes TWV. ATWV $= 1$ means a 100% accurate system (no FAs and no misses), whereas lower ATWV represents worse systems (with some wrong decisions). Systems with no output have ATWV $= 0$, but ATWV can even be negative (e.g. when no hits but lots of FAs are produced).

One drawback of TWV is that the cost of missing a hit depends on the number of true occurrences of the query in the search dataset. On the contrary, a false alarm is equally expensive for less as well as for widely occurring queries. The global TWV is averaged over each query's TWV, so each query has equal weight. That's why TWV "forces" to lower the threshold for less occurring queries. It is better to "pay a bit" for several false alarms than to "pay a lot" for one miss (especially for less occurring queries). This leads to a dependency of the threshold on the number of true query occurrences.

On the other hand, C_{nxe} is computed directly 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 (LLR). A perfect system would get $C_{nxe} = 0$ and a non-informative (but well calibrated) system would get $C_{nxe} = 1$, whereas $C_{nxe} > 1$ would indicate severe miscalibration of the log-likelihood ratio scores (see [21] for more details).

3.2. Overview of System Results

This year, there was a *Top-5* group of teams (NNI [9], BUT [8], SPL-IT [7], GTTS-EHU [14] and CUHK [10]), whose results were quite similar, while using a variety of different techniques. Figures 1 and 2 show the average results obtained by these participants, split across query subset and query language respectively, on the development and evaluation sets. In each plot, we show (actual) C_{nxe} and $1-$ATWV into a unified view. Note that although C_{nxe} and ATWV are quite different metrics (see below for a discussion) their trend is remarkably similar when averaged across teams.

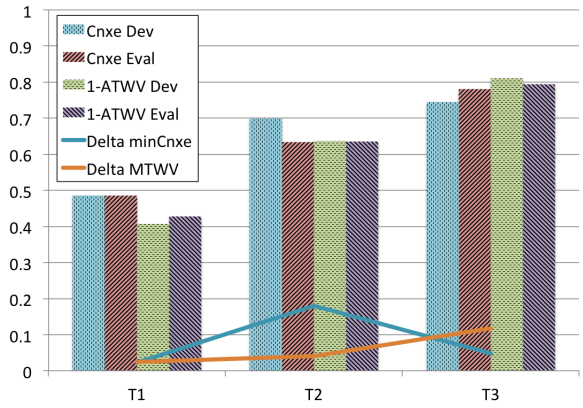

Fig. 1. Top-5 average results per query subset, using C_{nxe} (primary metric, lower is better), and $1 - $ ATWV (diagnostic metric, higher is better in ATWV).

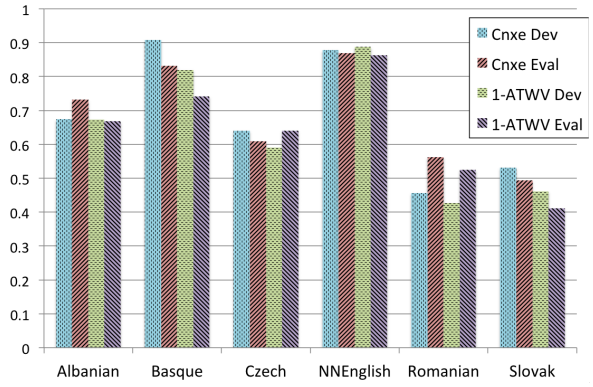

Fig. 2. Top-5 average results per query language (language that each query belongs to).

As expected, Figure 1 shows a decrease in performance from the easiest subset of queries T1 (involving exact matches) to T2 (allowing for small morphological variations) and T3 (allowing for both morphological and syntactic variations, as well as some filler content).

We investigated how participants optimized their systems to either C_{nxe} or ATWV (or did not try to optimize TWV at all), suggesting to also analyze system performance on minC_{nxe} and MTWV, which are somewhat "cheating" metrics. For these two metrics, we show the relative difference of performance on the development and evaluation sets in Figure 1 (Delta minC_{nxe} and Delta MTWV). On the T1 subset of queries, the difference is very small (i.e. systems are well tuned). On T2, the relative difference is larger (almost 20% on average) for minC_{nxe}, but remains small for MTWV. On T3, however, the relative difference is smaller for minC_{nxe} than for MTWV. It appears that confidence values, which are critical for C_{nxe} computation, are hard to compute accurately for T2 queries (which involve small morphological variations), but not so much for T3 queries (which involve stronger morpho-syntactic variations).

According to system descriptions, participants attempted to detect Type 3 matches in different ways, e.g. by integrating reorderings into DTW [7]. The retrieval of sub-strings in a symbolic approach was found to be particularly useful [9]. On the other hand, splitting queries in the middle and searching for halves was found not to help overall for symbolic approaches [8].

Per language analysis in Figure 2 shows that non-native English and Basque were the most difficult to match, probably due to challenging acoustic conditions in the search data and the low quality of pronunciations with regard to those of the queries (for non-native English). Also noticeable is the difference between C_{nxe} and ATWV in Basque and Slovak, and the difference between dev and eval queries in Romanian. The latter is most probably due to an imbalance between the difficulty of dev and eval queries, as the search corpus is fixed. Interestingly, good results were achieved on the Czech corpus in contrast to last year's evaluation, even though the search data were extracted from the same source [5]. This seems to be an effect of how queries were collected or recorded: this year Czech queries were spoken in isolation and recorded in an office environment, while last year they were cut from long (and noisy) telephone conversations.

3.3. Performance Metrics Comparison

One of the main novelties introduced in QUESST 2014 is the use of C_{nxe} as primary performance metric for the first time, the Term Weighted Value (TWV) playing the role of secondary metric. In this Section, we will give new insight into both metrics, through examples, and will briefly analyze how much they differ in evaluating the systems submitted to QUESST 2014.

As described above, C_{nxe} measures the goodness of score values, whereas TWV measures the goodness of discrete Yes/ No decisions. Therefore, two systems providing different scores but making the same decisions will have different C_{nxe} but the same TWV, as systems A and B in Figure 3. Are those systems equally good as TWV suggests, or not? The C_{nxe} metric tells us how much sensitive system decisions are to threshold variations. In the limit, C_{nxe} looks for target scores being $+\infty$ and non-target scores being $-\infty$. If that was the case, the system would always make the right decisions, no matter the threshold we chose. Note that C_{nxe} is an average over all the scores and does not depend on decisions. An extreme case is shown in Figure 3, where system C, despite making all right decisions (an thus, being TWV=1), is worse than system B in terms of C_{nxe}, just because the scores of system C are less separated one from another (on average) than those of system B (which, on the other hand, produces two wrong decisions, with TWV=0.3333).

System development involves the estimation of a score transformation and a global threshold in order to optimize for either TWV or C_{nxe} on the set of development queries. Optimizing for TWV means minimizing P_{fa} and P_{miss}, i.e. reducing the number of wrong decisions as much as possible. This could be attained by separating target and non-target scores as much as possible, i.e. by minimizing C_{nxe}, but there could be other ways to optimize for TWV, not requiring C_{nxe} minimization at all.

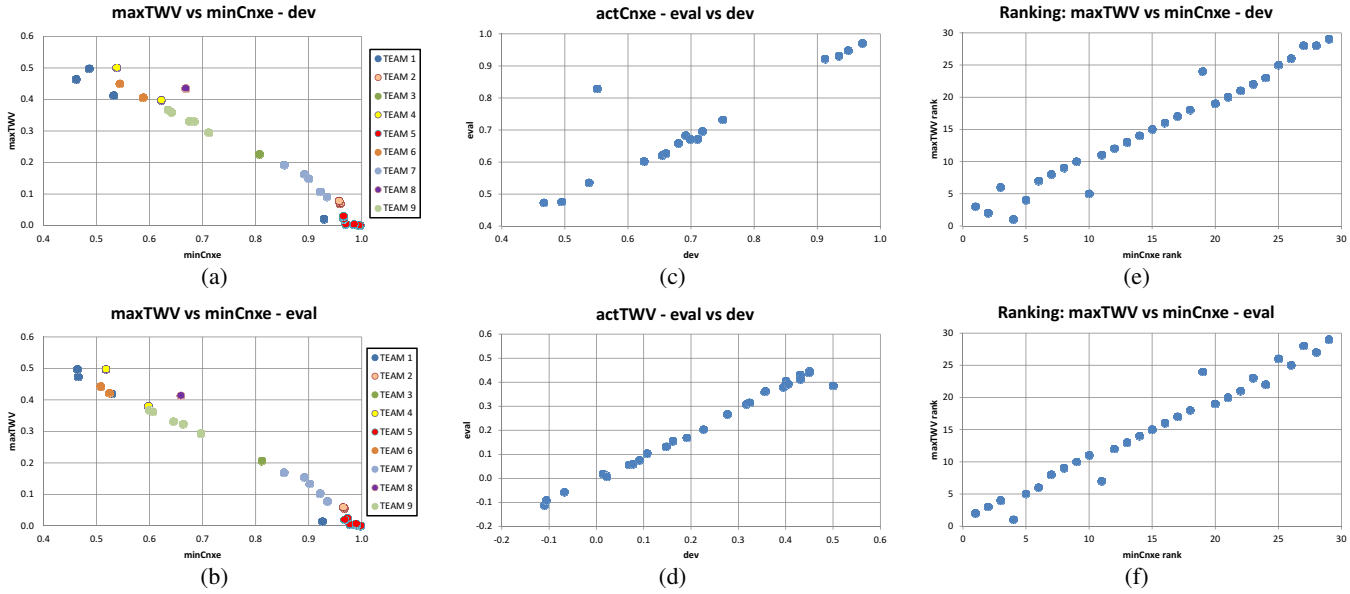

Fig. 4. Comparison between TWV and C_{nxe} performance in QUESST 2014: maxTWV vs minC_{nxe} (a,b); eval vs dev performance in terms of actC_{nxe} and actTWV (c,d); and maxTWV vs minC_{nxe} rankings (e,f).

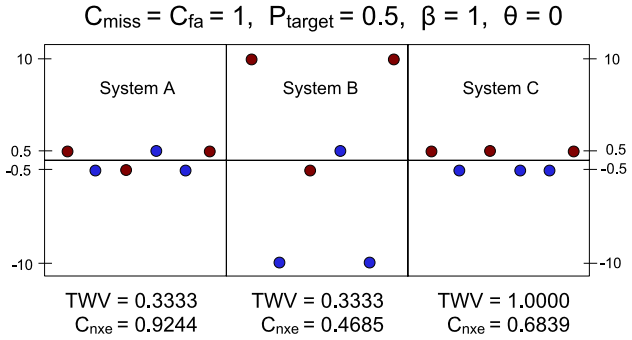

Fig. 3. Output scores of three example systems (A, B, C) for target (red) and non-target (blue) trials. The application parameters and the TWV and C_{nxe} values are shown too.

From a practical point of view, the main argument in favor of C_{nxe} is that "separating target and non-target scores as much as possible" has the side effect of allowing for suitable thresholds that provide low P_{fa} and P_{miss}, thus leading to good TWV values. On the contrary, optimizing for TWV also involves separating target and not target scores, but only to the extent that it is possible to define a threshold that provides low P_{fa} and P_{miss}, which could lead to comparatively poor C_{nxe} scores.

Besides, C_{nxe} provides more margin than TWV to make the right decisions in a region around the operating point for which the system is designed. Also, since C_{nxe} is computed from continuous scores, the risk of overfitting is lower than in the case of TWV, which is based on discrete decisions. Yet another argument in favor of C_{nxe} is that the threshold for making decisions is mathematically computed from the prior and costs that define the operating point [21]. It is not necessary to re-tune the system for a new operating point. Since C_{nxe} assumes that the scores are log-likelihood ratios, minimum expected cost decisions can be made. In fact, C_{nxe} just measures the goodness of those log-likelihood ratios. Theoretically, $1 - C_{nxe}$ tells us how much information is providing the system to make right decisions, with regard to a non-informative system.

But which metric, TWV or C_{nxe}, is preferable for system development? If the target application is known to work on a single operating point, then minimizing P_{fa} and P_{miss} around it would be enough, and thus TWV would be the best choice. If, instead, the target application must work on a wide range of operating points, then robustness to threshold variations is an important issue and C_{nxe} should be chosen.

Coming back to system results in QUESST 2014, the maxTWV and minC_{nxe} metrics seem to be highly correlated, the few outliers corresponding to systems that were very tightly optimized for TWV (see Figures 4.a and 4.b). As shown in Figures 4.c and 4.d, the actual TWV and C_{nxe} were both consistent across the sets of queries, except for one of those systems very tightly tuned on TWV, which was highly uncalibrated in terms of C_{nxe}. Finally, few and quantitatively small differences can be found between TWV and C_{nxe} rankings in both sets of queries (see Figures 4.e and 4.f).

4. CONCLUSIONS

This paper describes the "Query-by-Example Speech Search Task" (QUESST), held as part of the 2014 MediaEval benchmark campaign. The purpose of the evaluation was to perform language independent search on speech by using speech queries. The search database contains utterances from multiple languages, speakers and acoustic conditions, with no information neither on what is said nor in which language it is said. Three disjoint subsets of queries are defined, according to the most complex type of match (query occurrence) found in the search database, from exact matches (T1) to approximate matches allowing either for just small morphological variations (T2) or for both morphological and syntactic variations (T3). Besides describing the database, which has been made public for research purposes, we have discussed the main findings of the evaluation, by analyzing the average performance of the most competitive systems. We have also performed a comparative analysis of the two performance metrics used this year, namely the normalized cross entropy cost (C_{nxe}) and the Term-Weighted Value (TWV). We conclude that optimizing for minimum C_{nxe} results in more robust systems, which will also perform well for TWV, whereas the opposite is not always true.

5. REFERENCES

[1] MediaEval Benchmarking Initiative for Multimedia Evaluation, 2014, http://www.multimediaeval.org/mediaeval2014/.

[2] F. Metze, N. Rajput, X. Anguera, M. Davel, G. Gravier, C. van Heerden, G. V. Mantena, A. Muscariello, K. Prahallad, I. Szöke, and J. Tejedor, "The Spoken Web Search Task at MediaEval 2011," in *Proceedings of ICASSP*, Kyoto, Japan, March 25-30, 2012, pp. 5165–5168.

[3] F. Metze, X. Anguera, E. Barnard, M. Davel, and G. Gravier, "The Spoken Web Search Task at MediaEval 2012," in *Proceedings of ICASSP*, Vancouver, Canada, May 26-31, 2013, pp. 8121–8125.

[4] F. Metze, X. Anguera, E. Barnard, M. Davel, and G. Gravier, "Language independent search in MediaEval's Spoken Web Search task," *Computer Speech & Language*, vol. 28, no. 5, pp. 1066–1082, September 2014.

[5] X. Anguera, L.-J. Rodriguez-Fuentes, I. Szöke, A. Buzo, F. Metze, and M. Penagarikano, "Query-by-Example Spoken Term Detection on Multilingual Unconstrained Speech," in *Proceedings of Interspeech*, Singapore, September 14-18, 2014, pp. 2459–2463.

[6] X. Anguera, L.-J. Rodriguez-Fuentes, I. Szöke, A. Buzo, and F. Metze, "Query by Example Search on Speech at Mediaeval 2014," in *Working Notes Proceedings of the Mediaeval 2014 Workshop*, Barcelona, Spain, October 16-17, 2014.

[7] J. Proença, A. Veiga, and F. Perdigão, "The SPL-IT Query by Example Search on Speech system for MediaEval 2014," in *Working Notes Proceedings of the Mediaeval 2014 Workshop*, Barcelona, Spain, October 16-17, 2014.

[8] I. Szöke, M. Skácel, and L. Burget, "BUT QUESST 2014 system description," in *Working Notes Proceedings of the Mediaeval 2014 Workshop*, Barcelona, Spain, October 16-17, 2014.

[9] P. Yang, H. Xu, X. Xiao, L. Xie, C.-C. Leung, H. Chen, J. Yu, H. Lv, L. Wang, S. J. Leow, B. Ma, E. S. Chng, and H. Li, "The NNI Query-by-Example System for MediaEval 2014," in *Working Notes Proceedings of the Mediaeval 2014 Workshop*, Barcelona, Spain, October 16-17, 2014.

[10] H. Wang and T. Lee, "CUHK System for QUESST Task of MediaEval 2014," in *Working Notes Proceedings of the Mediaeval 2014 Workshop*, Barcelona, Spain, October 16-17, 2014.

[11] S. Kesiraju, G. Mantena, and K. Prahallad, "IIIT-H System for MediaEval 2014 QUESST," in *Working Notes Proceedings of the Mediaeval 2014 Workshop*, Barcelona, Spain, October 16-17, 2014.

[12] M. Calvo, M. Giménez, L.-F. Hurtado, E. Sanchis, and J. A. Gómez, "ELiRF at MediaEval 2014: Query by Example Search on Speech Task (QUESST)," in *Working Notes Proceedings of the Mediaeval 2014 Workshop*, Barcelona, Spain, October 16-17, 2014.

[13] J. Vavrek, P. Viszlay, M. Lojka, M. Pleva, and J. Juhar, "TUKE system for MediaEval 2014 QUESST task," in *Working Notes Proceedings of the Mediaeval 2014 Workshop*, Barcelona, Spain, October 16-17, 2014.

[14] L.-J. Rodriguez-Fuentes, A. Varona, M. Penagarikano, G. Bordel, and M. Diez, "GTTS-EHU Systems for QUESST at MediaEval 2014," in *Working Notes Proceedings of the Mediaeval 2014 Workshop*, Barcelona, Spain, October 16-17, 2014.

[15] A. Buzo, H. Cucu, and C. Burileanu, "SpeeD @ MediaEval 2014: Spoken Term Detection with Robust Multilingual Phone Recognition," in *Working Notes Proceedings of the Mediaeval 2014 Workshop*, Barcelona, Spain, October 16-17, 2014.

[16] J. Fiscus, J. Ajot, J. Garofolo, and G. Doddington, "Results of the 2006 Spoken Term Detection Evaluation," in *ACM SIGIR Workshop on Searching Spontaneous Conversational Speech*, Amsterdam, 2007.

[17] NIST Open Keyword Search 2014 Evaluation (OpenKWS14), *KWS14 Keyword Search Evaluation Plan*, December 2013, http://www.nist.gov/itl/iad/mig/upload/KWS14-evalplan-v11.pdf.

[18] Multilingual Database for the Query-by-Example Search on Speech Task (QUESST) at MediaEval 2014, Available online: http://speech.fit.vutbr.cz/files/quesst14Database.tgz.

[19] TED: Ideas worth spreading, http://www.ted.com/ and http://tedxtalks.ted.com/.

[20] National Institute of Standards and Technology (NIST), *The Spoken Term Detection (STD) 2006 Evaluation Plan*, September 2006, http://www.itl.nist.gov/iad/mig/tests/std/2006/.

[21] L.-J. Rodriguez-Fuentes and M. Penagarikano, "MediaEval 2013 Spoken Web Search Task: System Performance Measures," Tech. Rep., Software Technologies Working Group, University of the Basque Country UPV/EHU, May 2013, http://gtts.ehu.es/gtts/NT/fulltext/rodriguezmediaeval13.pdf.