THE SPOKEN WEB SEARCH TASK AT MEDIAEVAL 2012

Florian Metze[1], Xavier Anguera[2], Etienne Barnard[3], Marelie Davel[3], and Guillaume Gravier[4]

[1]Carnegie Mellon University; Pittsburgh, PA, USA (fmetze@cs.cmu.edu)
[2]Telefonica Research; Barcelona, Spain (xanguera@tid.es)
[3]North-West University; Vanderbijlpark, South Africa
({etienne.barnard|marelie.davel}@gmail.com)
[4]IRISA/ INRIA; Rennes, France (guig@irisa.fr)

ABSTRACT

In this paper, we describe the "Spoken Web Search" Task, which was held as part of the 2012 MediaEval benchmark evaluation campaign. The purpose of this task was to perform audio search with audio input in four languages, with very few resources being available. Continuing in the spirit of the 2011 Spoken Web Search Task, which used speech from four Indian languages, the 2012 data was taken from the LWAZI corpus, to provide even more diversity and allow for a task that will allow both zero resource "pattern matching" approaches and "speech recognition" based approaches to participate. In this paper, we summarize the results from several independent systems, developed by nine teams, analyze their performance, and provide directions for future research.

Index Terms— low-resource speech recognition, evaluation, pattern matching, spoken term detection

1. INTRODUCTION

MediaEval 2012's "Spoken Web Search" task involves searching *for* audio content, *within* audio content, *using* an audio content query.

The task therefore requires researchers to build a language-independent audio search systems so that, given a query, they should be able to find the appropriate audio file(s) and the (approximate) location of the query term within the audio file(s). The task is designed so that performing language identification, followed by standard speech-to-text is possible, but not the easiest solution as recognizers are typically not available in these languages. The evaluation was performed using standard NIST metrics for spoken term detection [1]. For comparison, participants could also search using the lexical form of the query, but dictionary entries for the search terms were not provided. We are not reporting results for this subtask in this paper. This way, the results shown here are therefore also relevant for the processing of languages or dialects for which written forms do not exist, which are part of the long tail of languages spoken in the world nowadays.

2. PRIOR AND RELATED WORK

This task was originally suggested by IBM Research India, and was initially run using data provided by this group, see [2], with the objective to be able to go beyond searching through meta-data only [3]. The 2012 evaluation is a continuation of the 2011 "Spoken Web Search" (SWS) Task [4]. The 2011 data was made available to 2012 participants as additional development data, although the scoring parameters had been adjusted to better reflect the intended use case. We will not report results on this data here.

Recently, there has been great interest in algorithms that allow rapid and robust development of speech technology for any language, particularly with respect to search, see for example [5] for an alternative approach. Today's technology has mostly been developed for transcription of English, with markedly lower performance on non-English languages, and still covering only a small subset of the world's languages. While most minority languages might not 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.

This evaluation attempts to provide an evaluation corpus and baseline for research on language-independent search and transcription of real-world speech data, with a special focus on low-resource languages, in order to provide a forum for original research ideas. The task is also suitable for young researchers to get started on speech technologies, and we are currently working on making some of the systems that were submitted to the 2012 evaluation available in the "Speech Recognition Virtual Kitchen" [6].

In this paper, we give an overview of the different approaches submitted to the evaluation, analyze the results, and summarize the findings of the evaluation workshop [7].

3. DESCRIPTION OF TASK AND DATA

The data used in the evaluation was divided into a development set, which was made available to participants about 16 weeks before the evaluation deadline, and an evaluation set, which the participants received about 8 weeks before the submission deadline. Both sets contained about 4 hours of audio.

By design, the 2012 "African" development data consists of 1580 audio files (395 per language), taken from the isiNdebele, Siswati, Tshivenda, and Xitsonga parts of the LWAZI corpus [8], and 100 example queries (25 per language) in these languages with overall similar characteristics to the 2011 "Indian" data [9, 4].

The evaluation data consists of 1660 audio files, and 100 queries; both data sets were selected from the LWAZI corpus to exhibit similar properties with respect to the frequency and distribution of the respective keyword sets. All these audio files were collected over a telephone channel, and provided as 8kHz/ 16bit WAV files.

As the focus is on language-independent search, no pronunciation dictionaries are provided by default, although these are available for contrastive experiments. Language labels are provided for the development data, but not for the evaluation data. The locations of the occurrences are provided, and participants received a scoring tool

based on the Actual Term Weighted Value (ATWV) metric [1], in which some of the internal parameters are modified to better represent a "useful" tradeoff between missed detections and false alarms for the "Spoken Web Search" scenario.

Systems were classified into two conditions: *restricted* (essentially the "zero resource" case where no external data is used) and *open*. Participants could not use any LWAZI corpus resources that had not been distributed for the evaluation, to avoid them inadvertently including evaluation data in development, they were free to include any other resource (i.e. existing phone recognizers, etc.) in their systems, as long as their use was documented.

4. SYSTEM DESCRIPTIONS

In the interest of brevity, we will describe in the following only an informative selection of systems submitted to the evaluation, leaving out some contrastive and diagnostic systems.

4.1. Open Systems

"Open" systems can use any resource available, as long as its use is documented, and should therefore be able to outperform "restricted" systems. Depending on the type of external resource chosen, the costs may be negligeable, even though systems are no longer "zero resource". Most open systems employ some sort of multi-lingual phonetic tokenizer to convert the audio data into a symbolic sequence, or frame-level posteriorgrams, which is then searched.

cuhk_phnrecgmmasm_p-fusionprf [10]

This system used five semi-supervised tokenizers [11] and two unsupervised tokenizers. The semi-supervised tokenizers used phoneme recognizers to convert the development audio data into posterior features, which were further transformed and modeled by a mixture of 256 Gaussians. Five semi-supervised tokenizers were built from Czech, Hungarian, Russian, English and Mandarin phoneme recognizers, which were all in the split temporal context network structure [12, 13]. The two unsupervised tokenizers were "GMM" and "ASM" (Acoustic Segment Modeling) [14], as described in Section 4.2. All these tokenizers were used to generate posteriorgrams, and Dynamic Time Warping (DTW) was applied for detection. To exploit the complementary information of all the tokenizers, a DTW matrix combination approach [11] was used. Pseudo relevance feedback (PRF) and score normalization were used as the back-end.

l2f_12_spch_p-phonetic4_fusion_mv [15]

The L2F SWS system consists of four sub-systems, which are each based on the AUDIMUS [16] Automatic Speech Recognition (ASR) system. They exploit four different language-dependent acoustic models trained for European Portuguese, Brazilian Portuguese, European Spanish, and American English.

A phone-loop grammar with phoneme minimum duration of three frames is used to obtain a phonetic transcription or tokenization for each query. In development experiments, no significant benefit was observed when using alternative n-best hypothesis for characterizing each query.

Spoken query search is based on Acoustic Keyword Spotting (AKWS) using the hybrid AUDIMUS WFST recognizer. The system combines four MLP outputs trained with 13 PLP features (plus deltas), 13 PLP-RASTA (plus deltas), Modulation SpectroGram features (28 static) and ETSI Advanced Front-end features (plus delta and delta-deltas). A search sliding window of 5 seconds with 2.5 seconds of time shift is used to process each file. An equally-likely 1-gram language model formed by the target query and a competing speech background model is used. The minimum duration for the background speech word is set to 250 ms.

To normalize the mean and variance of the resulting scores, a query dependent "Q-norm" normalization was applied. The four sub-systems were then fused using majority voting.

BUT_spch_p-akws-devterms [17]

The BUT AKWS systems extract 3-state phone posterior features or bottle-neck features encoding speech (queries and utterances) in low dimensional vectors. The feature extractor is the same as that presented in [18], and contains a Neural Network (NN) classifier with a hierarchical structure called bottle-neck universal context network. Energies from 15 Mel-scale critical bands, ranging from 64 Hz to 3800 Hz, are extracted and passed through a logarithm.

For keyword spotting, a phone recognizer on the above mentioned 5-layer phone posteriors bottle-neck NN was developed, without using any phoneme language model – only free phone loop. Using jack-knifing, this system achieved a phone accuracy of 66.0% on the development data.

Following [19], an HMM was built for each query using a dictionary, and the log likelihood ratio between a query model and a background model (free phone loop) was computed. The development forced alignment and graphemic transcription of queries to obtain reference pronunciation of each query were used. This system achieved an Maximum TWV of 0.737 and the Upper-bound Term Weighted Value (UBTWV) of 0.859. The UBTWV finds the best threshold for each query (maximizes the TWV per query) and then averages the scores. It can be considered as non-pooled MTWV and shows the room for calibration improvement, providing an (reference) "R-AKWS" upper bound.

The actual Acoustic Keyword Spotting is similar to the reference one (R-AKWS). Only no prior knowledge of queries (pronunciation) was used. The pronunciation of the queries was automatically generated using the above mentioned phone recognizer. After generating pronunciations (one per query), surrounding silence was stripped, achieving an MTWV of 0.453 and UBTWV 0.600 without any score calibration on the development data. Using calibration, MTWV improved to 0.493.

BUT_spch_g-DTW-devterms [17]

The DTW system was based on a simple template-matching algorithm on bottle-neck features, using cosine distance as a similarity function [20]. The primary ATWV metric is sensitive to the calibration of the detection scores. For each query, an ideal hard-decision threshold was computed, which was then predicted during test using a linear regression model based on a number of features, such as lengths, counts, scores, etc., which can be computed on the query.

arf_spch_p-asrDTWAlign_w15_a08_b04 [21]

The ASR-based system has two components, as suggested in the NIST 2006 evaluation campaign: the indexer and the searcher. The indexed data are the transcriptions of the content data set using a phone recognizer. The multilingual acoustic model is obtained by adapting an acoustic model trained on Romanian language (previously developed in [22]) with the development data from the envisaged languages. The phone mapping is made using the International Phonetic Alphabet and a confusion matrix.

The searching component is based on DTW. A sliding window, whose length is proportional to the query length, is used for localizing the term. The method is refined by penalizing short queries and large DTW match spread in order to reduce the false alarms. This choice is motivated by the fact that for shorter queries and larger DTW match spreads the probability to confuse to similar phone strings is higher. This method scales well, because once the data are indexed the searching problem is reduced to string comparison.

gtts_spch_p-phone_lattice [23]

As a first step, the open software BUT phone decoders for Czech, Hungarian and Russian [12] are applied to decode both the spoken queries and the audio documents. For each spoken query, the N phone decodings with the highest likelihoods are extracted from the phone lattice using SRILM's lattice tool. Then, the Lattice2Multigram (L2M) tool [24] by Dong Wang is applied. The Master Label Files files are then re-ranked, filtered and converted to STD files using a heuristic approach.

4.2. Restricted Systems

cuhk_spch_p-gmmasmprf [10]

This system used two unsupervised tokenizers trained from the development audio data. The first was a GMM tokenizer containing 1024 Gaussian mixtures. The input of the GMM tokenizer was 39-dimensional MFCC feature vectors, which had been processed with vocal tract length normalization (VTLN). The second is an ASM tokenizer [14], containing 256 ASM units. Each unit had 3 states with 16 Gaussian mixtures for each state. The input features for the ASM tokenizer were the same as those for the GMM tokenizer. Combination of these two tokenizers was performed by the DTW matrix combination approach [11]. PRF and score normalization were used as the back-end.

jhu_all_spch_p-rails [25]

The RAILS approach involves four primary processing stages: (1) each frame is mapped to a sortable bit signature using locality sensitive hashing (LSH), using the variant that preserves cosine distance; (2) sorted lists (the index) of the signatures in the search collection are constructed; (3) using the index, approximate nearest neighbor sets for each query frame are computed in logarithmic time, allowing the construction of a sparse similarity matrix between query and search collection; and (4) runs of similar frames are searched for with efficient sparse image processing techniques applied to the similarity matrix. During MediaEval system development, system performance was investigated as a function of only two RAILS parameters: the size of the neighborhood search beam, B, and the cosine similarity threshold for a frame-level comparison to make it into the sparse matrix, δ. The default values listed in [26] were used for all other RAILS parameters.

tid_sws2012_IRDTW [27]

First, standard MFCC-39 (10 ms scroll in 25 ms window) features are extracted from the acoustic data. Then, posterior probabilities from these features are obtained by using a posteriors background model [28] that resembles a GMM model but using EM and K-means iterations during training. Then, after labeling the energy level of each frame and eliminating those frames that have low energy a dynamic programming matching process is performed between queries and reference data. To do this, we used a novel method called IR-DTW [29] which uses information retrieval concepts to efficiently match the queries into the database. In addition, we also submitted a run using the DTW-based system we implemented for MediaEval 2011 [30]. After retrieving all matches for any of the two systems, a post-processing step involves an overlap detection to eliminate those matches that are highly overlapped with each other.

tum_spch_p-cdtw [31]

Cumulative Dynamic Time Warping (CDTW) method is a novel variant of the DTW algorithm for comparing two sequences. Several modifications are introduced compared to standard DTW: First, the distance function measuring the local match between points of both sequences is replaced with more general "step scores". These scores are calculated as the combinations of several "feature functions" associated to each pair of points, and they can depend on the local step taken (diagonal, horizontal or vertical).

The other main change compared to common DTW is that the "hard" maximum (or minimum) operation of the DTW with the *softmax* operation. As a consequence, the obtained alignment score between the two sequences takes into account all the possible alignment paths instead of only the optimal one. Furthermore, the alignment score is differentiable with respect to the feature weights, allowing for an learning of these values.

MFCC+delta+acceleration descriptors were computed using HTK, and CMN/CVN was applied. The alignment features correspond to several functionals of these descriptors. For the retrieval task, a heuristic search for matching sequence candidates is first performed and the CDTW score is used as a decision score.

tuke_spch_p-dtwsvm [32]

The substantial step in comparing audio content of two segments is to extract parameters (coefficients) that capture temporal and spectral characteristics of the audio signal. We have tried several combinations of features such as MFCCs, ZCR and MPEG-7 low level descriptors (ASS, ASC, ASF, ASE) and chose the one with the average minimum-cost alignment (avgMCA), by using DTW algorithm, between selected queries and corresponding terms within utterances. The optimal results were achieved by using first 12 MFCCs features and zero energy coefficient (the avgMCA was about 250.1).

The main functionality of this proposed search algorithm lies in comparing two audio segments, with the same length, by using DTW algorithm and misclassification rate of Support Vector Machine (SVM) classifier. The first segment represents the searching query and the second one refers to the audio segment of each utterance with the same length as the query segment. Lin et al. [33] have used a similar procedure for detecting speaker changes, based on a new SVM training misclassification rate.

5. RESULTS AND ANALYSIS

Figure 1 shows the results of the primary system that each participant submitted on development data, which participants used to tune their approaches. Figure 2 shows the corresponding results on the unseen evaluation data.

Table 1 lists the results for these systems on both development and evaluation data, as scored by the organizers. This table lists the primary *Actual* Term Weighted Value metric achieved by the systems, while Figure 1 and Figure 2 show the *Maximum* TWV, so most systems were well tuned.

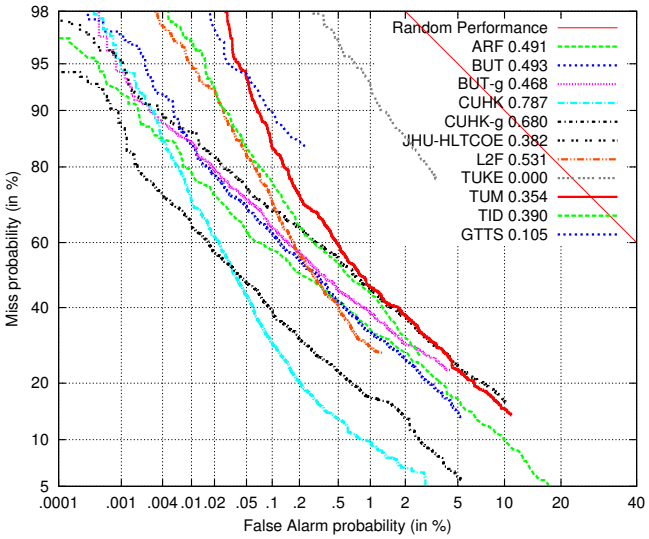

Fig. 1. DET (Detection Error Tradeoff) plots and MTWV (Maximum TWV) results for development terms on development data.

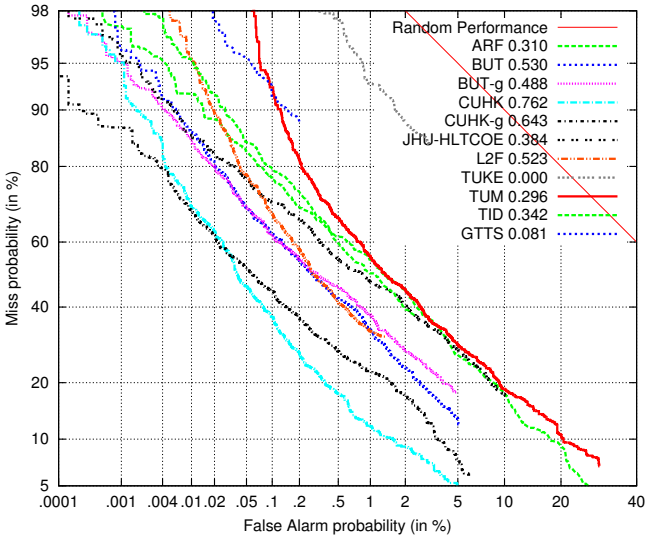

Fig. 2. DET plots and MTWV results for evaluation terms on evaluation data.

Table 1. Results (Actual TWV) for selected SWS 2012 systems.

System	Type	Dev	Eval
cuhk_phnrecgmmasm_p-fusionprf (CUHK)	open	0.782	0.743
cuhk_spch_p-gmmasmprf (CUHK-g)	restricted	0.678	0.635
l2f_12_spch_p-phonetic4_fusion_mv	open	0.531	0.520
BUT_spch_p-akws-devterms (BUT)	open	0.488	0.492
BUT_spch_g-DTW-devterms (BUT-g)	open	0.443	0.448
jhu_all_spch_p-rails (JHU-HLTCOE)	restricted	0.381	0.369
tid_sws2012_IRDTW	restricted	0.387	0.330
tum_spch_p-cdtw	restricted	0.263	0.290
arf_spch_p-asrDTWAlign_w15_a08_b04	open	0.411	0.245
gtts_spch_p-phone_lattice	open	0.098	0.081
tuke_spch_p-dtwsvm	restricted	0	0

It is interesting to note that under the given conditions, the zero-knowledge ("restricted") approaches could perform quite similarly to "open" (typically model-based) approaches, which typically rely on the availability of matching data from other languages. The difference is about 0.1 for the two CUHK systems, but 0.05 for the two BUT systems; the BUT DTW-based system used a tokenizer based on phone labels, so it was also classified as an "open" system. One of the reasons for the good performance of the CUHK systems may be the careful application of VTLN, which none of the other participants attempted, and the fusion of similarity matrices obtained using several different front-ends. The BUT experiments show that not having a lexicon available for query terms greatly impacts the performance of the AKWS system.

While not discussed in detail at the workshop, participants reported that this year's systems beat the 2011 systems on the 2011 data set, which used a different cost function for deletions versus insertions. To keep the load light for participants, no further measurements were required, so we cannot report on the complexity (both for building and at runtime) of the approaches, but participants reported that most approaches were quite lightweight. Also, participants used a wide range of resources and techniques, including the Brno phone recognizers [13] and neural network front-ends.

As some of the submitted systems consisted themselves already of several components, the organizers fused various system outputs in the "open" and "restricted" conditions using the CombMNZ algorithm as a post-evaluation experiment. The fused performance would beat the performance of the best system involved in the fusion, but not by much (ca. 0.03) for cases involving the CUHK systems.

6. CONCLUSIONS AND OUTLOOK

The results of the second SWS task at MediaEval shows progress on low-resource spoken term detection, and discussion at the workshop has already sparked a number of new investigations. The task investigates the fundamentals of how audio content can be made searchable, without having to develop a dedicated speech recognizer and dialog system, which is still a significant effort, particularly for resource-scarce settings.

With respect to the amount of in-domain data available, this task is even harder than the research goals proposed by, for example, IARPA's Babel [34] program, yet results have been achieved that appear useful in the context of the "Spoken Web" task, which is targeted primarily at communities that currently do not have access to Internet. Many target users have low literacy skills, and many speak in languages for which fully developed speech recognition systems will not exist even for years to come.

This work therefore presents an approach that can enable voice interaction in resource scarce settings, with all associated benefits, and also presents a lightweight test-bed in which ideas in many fields relevant to speech processing can be evaluated easily, which we intend to develop further in the future, using LWAZI data (which is already available to the public), or other similar corpora.

7. ACKNOWLEDGMENTS

The authors would like to acknowledge the MediaEval Multimedia Benchmark [35], and IBM Research India for providing the inspiration which sparked the first "Spoken Web Search" Task at MediaEval 2011. The organizers would also like to thank Martha Larson from TU Delft for organizing this event, and the participants for their hard work on this challenging evaluation.

8. REFERENCES

[1] J. Fiscus, J. Ajot, J. Garofolo, and G. Doddington, "Results of the 2006 spoken term detection evaluation," in *Proc. SSCS*, Amsterdam; Netherlands, 2007.

[2] A. Kumar, N. Rajput, D. Chakraborty, S. K. Agarwal, and A. A. Nanavati, "WWTW: The world wide telecom web," in *NSDR 2007 (SIGCOMM workshop)*, Kyoto, Japan, Aug. 2007.

[3] M. Diao, S. Mukherjea, N. Rajput, and K. Srivastava, "Faceted search and browsing of audio content on spoken web," in *Proc. CIKM*, 2010.

[4] N. Rajput and F. Metze, "Spoken web search," In *Proc. MediaEval 2011* [36].

[5] D. Can, E. Cooper, A. Ghoshal, M. Jansche, S. Khudanpur, B. Ramabhadran, M. Riley, M. Saraçlar, A. Sethy, M. Ulinski, and C. White, "Web derived pronunciations for spoken term detection," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 2009, SIGIR '09, ACM.

[6] F. Metze and E. Fosler-Lussier, "The speech recognition virtual kitchen: An initial prototype," in *Proc. INTERSPEECH*, Portland, OR, Sept. 2012, ISCA.

[7] F. Metze, E. Barnard, M. Davel, C. van Heerden, X. Anguera, G. Gravier, and N. Rajput, "The spoken web search task," In *Proc. MediaEval Workshop* [35], http://www.multimediaeval.org/mediaeval2012/sws2012/.

[8] E. Barnard, M. H. Davel, and C. van Heerden, "ASR corpus design for resource-scarce languages," in *Proc. INTERSPEECH*, Brighton; UK, Sept. 2009, pp. 2847–2850, ISCA.

[9] 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 *Proc. ICASSP*, Kyoto; Japan, Mar. 2012, IEEE.

[10] H. Wang and T. Lee, "CUHK system for the spoken web search task at MediaEval 2012," In *Proc. MediaEval 2012* [35].

[11] H. Wang, T. Lee, C.-C. Leung, B. Ma, and H. Li, "Using parallel tokenizers with DTW matrix combination for low-resource spoken term detection," in *Proc. ICASSP*. IEEE, May 2013.

[12] P. Schwarz, *Phoneme Recognition based on Long Temporal Context*, Ph.D. thesis, Brno University of Technology, 2009.

[13] "Phoneme recognizer," http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context.

[14] H. Wang, C. Leung, T. Lee, B. Ma, and H. Li, "An acoustic segment modeling approach to query-by-example spoken term detection," in *Proc. ICASSP*, Kyoto; Japan, Mar. 2012, IEEE.

[15] A. Abad and R. F. Astudillo, "The L2F spoken web search system for MediaEval 2012," In *Proc. MediaEval 2012* [35].

[16] H. Meinedo, A. Abad, T. Pellegrini, I. Trancoso, and J. Neto, "The L2F broadcast news speech recognition system," in *Proc. Fala*, Nov. 2010.

[17] I. Szöke, M. Fapšo, and K. Veselý, "BUT 2012 approaches for spoken web search - MediaEval 2012," In *Proc. MediaEval 2012* [35].

[18] F. Grézl and M. Karafiát, "Hierarchical neural net architectures for feature extraction in ASR," in *Proc. INTERSPEECH*, Makuhari; Japan, Sept. 2010, ISCA.

[19] I. Szöke, P. Schwarz, L. Burget, M. Karafiát, P. Matějka, and J. Černocký, "Phoneme based acoustics keyword spotting in informal continuous speech," *Lecture Notes in Computer Science*, vol. 8, no. 3658, Aug. 2005.

[20] J. Tejedor, I. Szöke, and M. Fapšo, "Novel methods for query selection and query combination in query-by-example spoken term detection," in *Proc. Searching Spontaneous Conversational Speech (SSCS)*, Florence; Italy, 2010.

[21] A. Buzo, H. Cucu, M. Safta, B. Ionescu, and C. Burileanu, "ARF @ MediaEval 2012: A romanian ASR-based approach to spoken term detection," In *Proc. MediaEval 2012* [35].

[22] H. Cucu, L. Besacier, C. Burileanu, and A. Buzo, "Investigating the role of machine translated text in ASR domain adaptation: Unsupervised and semi-supervised methods," in *Proc. ASRU*, Honolulu, HI; USA, Dec. 2011, IEEE.

[23] A. Varona, M. Penagarikano, L. J. Rodriguez-Fuentes, G. Bordel, and M. Diez, "GTTS system for the spoken web search task at MediaEval 2012," In *Proc. MediaEval 2012* [35].

[24] D. Wang, S. King, and J. Frankel, "Stochastic pronunciation modelling for out-of-vocabulary spoken term detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 9, no. 4, 2011, http://homepages.inf.ed.ac.uk/v1dwang2/public/tools/index.html.

[25] A. Jansen, B. van Durme, and P. Clark, "The JHU-HLTCOE spoken web search system for MediaEval 2012," In *Proc. MediaEval 2012* [35].

[26] A. Jansen and B. V. Durme, "Indexing raw acoustic features for scalable zero resource search," in *Proc. INTERSPEECH*, Portland, OR; USA, Sept. 2012, ISCA.

[27] X. Anguera, "Telefonica research system for the spoken web search task at MediaEval 2012," In *Proc. MediaEval 2012* [35].

[28] X. Anguera, "Speaker independent discriminant feature extraction for acoustic pattern-matching," in *Proc. ICASSP*, Kyoto; Japan, Mar. 2012, IEEE.

[29] G. Mantena and X. Anguera, "Speed improvements to information retrieval-based dynamic time warping using hierarchical k-means clustering," in *Proc. ICASSP*. IEEE, May 2013.

[30] X. Anguera, "Telefonica system for the spoken web search task at MediaEval 2011," In *Proc. MediaEval 2011* [36].

[31] C. Joder, F. Weninger, M. Wöllmer, and B. Schuller, "The TUM cumulative DTW approach for the MediaEval 2012 spoken web search task," In *Proc. MediaEval 2012* [35].

[32] J. Vavrek, M. Pleva, and J. Juhár, "TUKE MediaEval 2012: Spoken web search using DTW and unsupervised SVM," In *Proc. MediaEval 2012* [35].

[33] P.-C. Lin, J.-C. Wang, J.-F. Wang, and H.-C. Sung, "Unsupervised speaker change detection using SVM training misclassification rate," *IEEE Trans. Comput.*, vol. 56, no. 9, pp. 1212–1244, Sept. 2007.

[34] Intelligence Advanced Research Projects Activity, "IARPA-BAA-11-02," http://www.iarpa.gov/solicitations_babel.html, 2011.

[35] *MediaEval 2012 Workshop*, Pisa, Italy, Oct. 2012. http://www.multimediaeval.org/mediaeval2012/.

[36] *MediaEval 2011 Workshop*, Pisa, Italy, Sept. 2011. http://www.multimediaeval.org/mediaeval2011/.