THE SPOKEN WEB SEARCH TASK AT MEDIAEVAL 2011

Florian Metze[1], *Nitendra Rajput*[2], *Xavier Anguera*[3], *Marelie Davel*[4], *Guillaume Gravier*[6], *Charl van Heerden*[4], *Gautam V. Mantena*[7], *Armando Muscariello*[6], *Kishore Prahallad*[7], *Igor Szöke*[5], *and Javier Tejedor*[8]

[1]Carnegie Mellon University; Pittsburgh, PA, USA
[2]IBM Research; New Delhi, India
[3]Telefonica Research; Barcelona, Spain
[4]North-West University, Vanderbijlpark; South Africa
[5]Speech@FIT, Brno University of Technology; Czech Republic
[6]IRISA/ INRIA; Rennes, France
[7]International Institute of Information Technology; Hyderabad, India
[8]HCTLab, Universidad Autónoma de Madrid; Spain
fmetze@cs.cmu.edu, rnitendra@in.ibm.com

ABSTRACT

In this paper, we describe the "Spoken Web Search" Task, which was held as part of the 2011 MediaEval benchmark campaign. The purpose of this task was to perform audio search with audio input in four languages, with very few resources being available in each language. The data was taken from "spoken web" material collected over mobile phone connections by IBM India. We present results from several independent systems, developed by five teams and using different approaches, compare them, and provide analysis and directions for future research.

Index Terms— low-resource speech recognition, evaluation, spoken web, audio search, spoken term detection

1. INTRODUCTION

The "Spoken Web Search" task of MediaEval 2011 involves searching *for* audio content, *within* audio content, *using* an audio content query. By design, the data consisted of only 700 utterances in Telephony quality from four Indian languages (English, Gujarati, Hindi, Telugu), without language labels. The task therefore required researchers to build a language-independent audio search system so that, given a query, it should be able to find the appropriate audio file(s) and the (approximate) location of query term within the audio file(s). Performing language identification, followed by standard speech-to-text is not appropriate, because recognizers are typically not available in these languages. 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, and we are not reporting results here. This work is the first in which results under this condition have been reported, and encompasses languages or dialects with no written form.

This task has been suggested by IBM Research India, and is using data provided by this group, see [2], to be able to go bayond searching through the meta-data related to the audio content only [3].

Recently, there has been a great interest in algorithms that allow rapid and robust development of speech technology for any language, particularly with respect to search, see for example [4]. Today's technology was mostly 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.

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.

In this paper, we will give an overview of the different approaches submitted to the evaluation [5, 6, 7, 8, 9], analyze the results, and summarize the findings of the evaluation workshop [10].

2. DESCRIPTION OF TASK AND DATA

Participants were provided with a dataset that has been kindly made available by the Spoken Web team at IBM Research India [2]. The audio content is spontaneous speech that has been recorded using commonly available land-line and mobile phone equipment in a live setting by low-literate users. While most of the audio content is related to farming practices, there are other domains as well. Data was collected from the following domains: (1) Sugarcane information by farmers (North India, Hindi language, 3000 users), (2) Village portal by villagers (South India, Telugu language, 6500 users), (3) Farming knowledge portal (West India, Gujarati language, 175 users), (4) Mixed content, e.g. job portal, event agenda (South and North India, English language, 80 users).

Table 1 provides details of the selected data. The labels identifying the language were intentionally not provided either in the development or the evaluation dataset.

The development set contains 400 utterances (100 per language) or "documents", and 64 queries (16 per language), all provided as digital recordings in 8kHz/ 16bit quality. For each query (and document on the development data), Romanized lexical transcriptions in UTF-8 encoding were also provided. The transcriptions had been generated by native speakers. For each development document, up to n matching queries were provided to participants, but not the exact location of the match within the document. A "match" is defined by the word transcription of the query appearing identically in the document. Sequences of one to three words were included as queries. they were selected to include typical target phrases, names, etc., which occur with a minimum frequency.

Category	# Utts	Total (h)	Average (sec)
Dev Documents	400	2:02:22	18.3
Dev Queries	64	0:01:19	1.2
Eval Documents	200	0:47:04	14.1
Eval Queries	36	0:00:58	1.6
Total	700	2:51:42	14.7

Table 1. Development (Dev) and evaluation (Eval) data used for the "Spoken Web Search Task" at MediaEval 2011.

Evaluation data consists of 200 utterances (50 per language), and 36 queries (9 per language), selected using the same criteria. Participants were allowed to use any additional resources they might have available, as long as their use is documented.

Participants received development audio data (documents and queries) as well as matches between queries and documents, and had five weeks to develop systems, before they also received the evaluation audio data. Results were due another five weeks later. There was no overlap between development and evaluation speakers, samples, and queries. Participants also attempted to detect occurrences of development queries on evaluation data, as well as evaluation queries on development data. The purpose of requiring these two additional, contrastive conditions was to see how critical tuning is for the different approaches: participants knew their performance for "dev queries" on "dev documents", so for evaluation we will evaluate the performance of unseen "eval queries" on previously known "dev documents" (which could have been used for unsupervised adaptation, etc.), known queries (for which good classifiers could have been developed) on unseen data, and unseen queries on unseen utterances (the primary condition). No group however achieved reasonable performance on the dev-eval and eval-dev conditions, and we assume this is due to the difficulty in choosing good parameters for an overall low number of positive events, and acoustic mismatch, so we will not discuss these results in the following.

Data was provided as a "term-list" XML file, in which the "term-id" corresponds to the file-name of the audio query. This information was distributed along with a modified version of the NIST 2006 Spoken Term Detection (STD) evaluation scoring software [1]. The primary evaluation metric was ATWV (Average Term Weighted Value), computed with the default values for the 2006 STD evaluation with respect to density of search terms, etc., but with "similarity" and "find" thresholds set to 20 s.

3. SYSTEM DESCRIPTIONS

In the following, we describe a selection of systems submitted to the evaluation. More systems were submitted, but the following ones provide the greatest insight and variety of approaches.

Broadly speaking, IRISA and TID submitted "zero-knowledge" approaches trained only on the available data, while BUT-HCTLabs and MUST submitted phone-based systems, which leveraged additional information. IIIT submitted an articulatory-feature-based approach, which also leveraged additional audio data.

3.1. GMM/HMM Term Modeling – BUT-HCTLabs

This approach is inspired by filler model-based acoustic keyword spotting, with a standard Viterbi-based decoder slightly modified to compute a likelihood ratio [11]. However, instead of the typical representation of both query model consisting of the corresponding phone transcription of the query term and filler/background model consisting of a free phone loop, we stuck with an acoustic representation of both the query term and the filler/background model, to maintain the language-independency, as follows: (1) the query model is represented with a Gaussian Mixture Model/Hidden Markov Model (GMM/HMM) whose number of states is 3 times the number of phones according to the phone recognition with 1 GMM component each, (2) the background model is a GMM/HMM with 1-state modeled with 10-GMM components. Queries represented by a single phone have been modeled with 6 states as if the query contained 2 phones to prevent the system from generating many false alarms for those queries. We used the number of phones output by a Slovak recognizer to derive the number of states of each query model due to its best performance in terms of the Upper-bound Term Weighted Value metric (UBTWV) [12] among Czech, English, Hungarian, Levantine, Polish, Russian, and Slovak. The score assigned by the acoustic keyword spotter to each detection is the likelihood ratio divided by the length of the detection. To deal with the score calibration and some problematic query length, detections were post-processed by 'Filtering" according to a length difference from an "average length" criterion. The average length of a query is calculated as the average length of speech (phones) across all the phone recognizers, except the Polish one due to its worse performance in the development data. Next, each detection is re-scored: the detection score remains the same if the detection length is longer than 80% and shorter than 140% of the average query length. Otherwise the score is lowered the shorter/longer the detection is according to the original query.

Features used for background and query modeling were obtained as follows: (1) the 3-state log-phone posteriors obtained by concatenating all the 3-state phone posteriors according to each feature extractor (corresponding to the recognizers of Czech, English, Hungarian, Levantine, Polish, Russian and Slovak languages) are applied a Karhunen-Loeve transform (KLT) for each individual language, (2) we keep the features that explain up to 95% of the variance after KLT for each individual language, (3) we build a 152-dimensional feature super-vector, from them. The KLT statistics have been computed from the development data provided by the task organizers and next applied on the evaluation data. The 3-state phone posterior estimator [13] contains a Neural Network (NN) classifier with a hierarchical structure called *bottle-neck universal context network*. It consists of a context network, trained as a 5-layer NN, and a merger that employs 5 context net outputs. The nets use $\approx$ 40 phones, $\approx$ 1300 nodes in the hidden layer, $\approx$ 480 nodes of the hidden layer in the merger net, $\approx$ 120 nodes in the posterior output layer, see [13].

3.2. Articulatory Features and Sliding DTW – IIIT

The primary motivation for this approach is to have speech specific features rather than language specific features like phone models. The advantage is that the articulatory phonetic units (selected well) could represent a broader set of languages. This would enable us to build articulatory units from one language and use it for other languages. We used 15 articulatory categories.

Audio content is decoded into their corresponding articulatory units using HMM models with 64 Gaussian mixture models using HTK. The models were built using 15 hours of telephone Telugu data, consisting of 200 speakers. Using the decoded articulatory phonetic output, tri-grams were used for indexing. The audio query was also decoded and the audio segments were selected, if there was a match in any of the tri-grams. Let t_{start} and t_{end} be the start and the end time-stamps for the tri-gram in the audio content that matches with one of the tri-grams from the audio query. Then the likely segment from the audio content would be between

$(t_{start} - audio\, query\, length)$ and $(t_{end} + audio\, query\, length)$.

These time stamps would provide the audio segments that are likely to contain the audio query. Sliding Dynamic Time Warping (DTW) search was applied on these audio segments to obtain the appropriate time stamps for the query.We propose an approach where we consider an audio content segment of length twice the length of the audio query, and a DTW is performed. After a segment has been compared the window is moved by one feature shift and DTW search is computed again. MFCC features, with window length 20 msec and 10 msec window shift, have been used to represent the speech signal. Consider an audio content segment S and an audio query Q. Construct a substitution matrix M of size qxs_q where q is the size of Q and $s_q = 2q$. We also define $M[i, j]$ as the node measuring the optimal alignment of the segments $Q[1 : i]$ and $S[1 : j]$.

During DTW search, at some instants $M[q, j](j <= s_q)$ will be reached. Then the time instants from column j to column s_q are the possible end points for the audio segment. Euclidean distance measure has been used to calculate the costs for the matrix M. The scores corresponding to all the possible end points are considered for k-means clustering. For $k = 3$, mean scores are calculated. The segment with the lowest score among segments overlapping at least 70% is selected.

3.3. Template Matching – IRISA

This system purely relies on pattern matching, exploiting different pattern comparison approaches to deal with variability in speech. The system operates at the acoustic level, with limited prior knowledge eventually embedded in posteriorgrams. As in most (if not all) pattern matching approaches, a segmental variant of DTW is used to efficiently search for the query in each document. Candidate hits are further evaluated with self-similarity matrix comparison. Details can be found in [14].

3.4. Pattern Matching with DGMM Posteriorgrams – TID

This system is based exclusively on a pattern-matching approach, which is able to perform a query-by-example search with no prior knowledge of the acoustics or language being spoken, computed on the non-silence part of the data. For the main submission we construct a Discriminative Gaussian Mixture Model (DGMM) [15] and store the Gaussian posterior probabilities (normalized to sum to 1) as features. Differently from standard GMM-posteriors, in DGMM modeling after the standard Enhanced Max Margin Learning (EMML) GMM training step we perform a hard assignment of each frame to their most likely Gaussian and retrain the Gaussian's mean and variance to optimally model these frames. This last step tries to solve a problem that EMML training has, which focuses on optimizing the Gaussian parameters to maximize the overall likelihood of the model on the input data, but not on discriminating between the different sounds in it. By performing the last assignment and retraining step we push Gaussians apart from each other to better model individual groups of frames depending on their location and density. This results in Gaussians with much less overlap, thus obtaining more discriminative posterior probability feature vectors. For this evaluation, only the development data from the SWS task was used for training.

In the comparison step, given two sequences, X and Y of posterior probabilities, respectively obtained from the query and any given phone recording, we compare them using a DTW-like algorithm. The standard DTW algorithm returns the optimum alignment between any two sequences by finding the optimum path between their start $(0, 0)$ and end (x_{end}, y_{end}) points. In our case we constrain the query signal to match between start and end, but we allow the phone recording to start its alignment at any position $(0, y)$ and finish its alignment whenever the dynamic programming algorithm reaches $x = x_{end}$. Although we do not set any global constraints, the local constraints are set so that at maximum 2-times or $\frac{1}{2}$-times warping is allowed. In addition, at every matching step we normalize the scores by the length of the optimum path up to that point, slightly favoring diagonal matches.

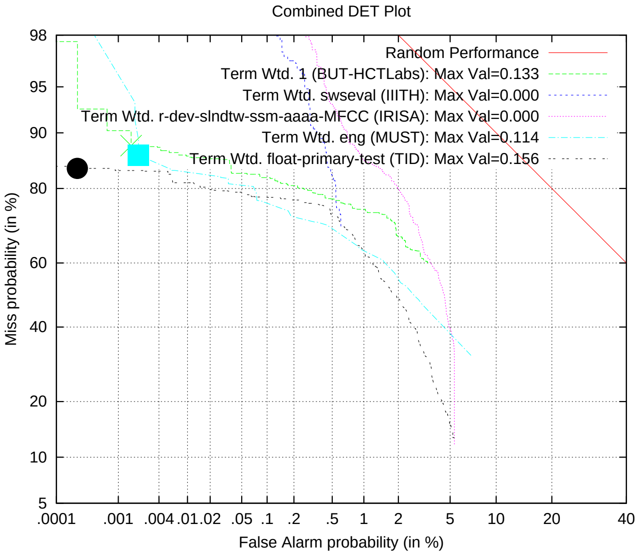

Fig. 1. Results (ATWV) on development data.

3.5. Phone Recognition – MUST

Aiming at both speaker independence and robustness with respect to recognition errors in the spoken queries, we implemented a phone-based system. The main dataset used for acoustic modeling was 60 hours of spontaneous conversations in colloquial Hindi. There are 996 native Hindi speakers and all conversations range between 1 and 4 minutes in duration. All conversations are transcribed and a basic pronunciation dictionary is provided.

A standard HMM-based recognizer was constructed using the Hindi data. The list of mono-phones was reduced from 62 to 21 units in order to work with a small set of broad but reliable classes, appropriate to the later scoring tasks. The speech data and transcriptions provided were cleaned using an aggressive form of garbage modeling, and the resulting data used to Maximum-A-Posteriori-adapt the Hindi acoustic models to the task domain (and languages).

Unconstrained phone recognition of both the query terms and the content audio is employed to represent these recordings as phone strings. A dynamic-programming (DP) approach with a linguistically motivated confusion matrix then finds regions in the content phone strings that correspond closely to one or more query strings. The resulting DP score (normalized by phone length) is used as confidence measure.

4. RESULTS AND ANALYSIS

Figure 1 shows the results of the five approaches described above on the development data, which participants could use to develop and

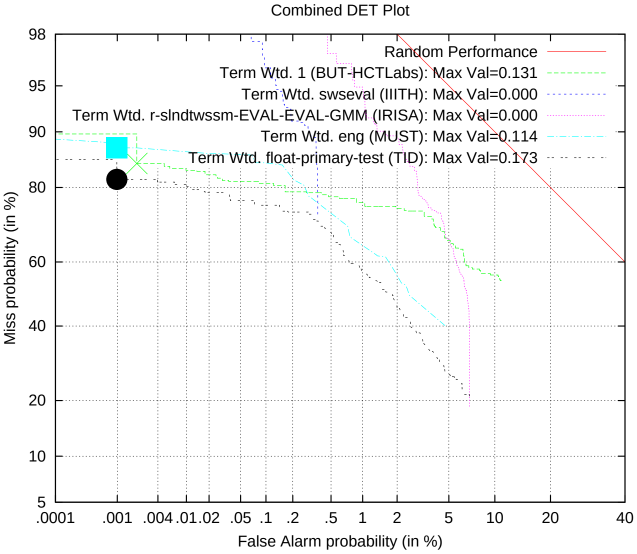

Fig. 2. Results (ATWV) on evaluation data.

tune their approaches. Using the provided metric, three approaches achieved maximal ATWVs of 0.1 or greater, detecting about 15% of events, with very low false alarm probabilities, as required by the parameters chosen to compute ATWV.

Figure 2 shows the corresponding results on the unseen evaluation data. The same three approaches could be run successfully on the unseen evaluation data, and the choice of decision thresholds was remarkably stable, given the little amount of available data. The "TID" query-by-example approach generalizes best to unseen data, while the phone-based "MUST" approach achieves an identical performance on the evaluation data as on the development data. The "BUT-HCTLabs" HMM/GMM approach also generalizes well.

It is interesting to note that under the given conditions, the zero-knowledge approaches could perform very similarly to the phone-based approaches, which relied on the availability of matching data from other languages. Follow-up work on larger datasets will be required to investigate the scalability of these systems.

5. OUTLOOK

These initial results on a very low-resource spoken term detection task show promising results, which will be explored further and improved upon in future work. It is interesting to note that very diverse approaches could achieve very similar results, and future work should include more evaluation criteria, such as amount of external data used, processing time(s), etc., which were deliberately left unrestricted in this evaluation, to encourage participation.

With respect to the amount of data available, this evaluation was even harder than the research goals proposed by for example IARPA's Babel [4] 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 won't exist even for years to come. Yet, access to highly variable information is critical for their development.

The organizers and participants are currently working to prepare more and varied data for future evaluations in a similar style, for example the Lwazi corpus [16]. We will attempt to make this, and other, future evaluation corpora available to a wider audience, in order to promote insight and progress on making speech technology available independent of the speaker's language.

6. ACKNOWLEDGMENTS

The authors would like to acknowledge the MediaEval Multimedia Benchmark (http://www.multimediaeval.org/mediaeval2011/index.html), and IBM Research India for collecting and providing the audio data and references used in this research. 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.

7. 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] "IARPA-BAA-11-02," http://www.iarpa.gov/solicitations_babel.html.

[5] X. Anguera, "Telefonica system for the spoken web search task at MediaEval 2011," In *Proc.* [17].

[6] E. Barnard, M. Davel, C. van Heeren, N. Kleynhans, and K. Bali, "Phone recognition for spoken web search," In *Proc.* [17].

[7] G. V. Mantena, B. Babu, and K. Prahallad, "SWS Task: Articulatory phonetic units and sliding DTW," In *Proc.* [17].

[8] I. Szöke, J. Tejedor, M. Fapšo, and J. Colás, "BUT-HCTLab approaches for spoken web search," In *Proc.* [17].

[9] A. Muscariello and G. Gravier, "IRISA MediaEval 2011 spoken web search system," In *Proc.* [17].

[10] N. Rajput and F. Metze, "Spoken web search," In *Proc.* [17].

[11] I. Szöke, P. Schwarz, P. Matějka, L. Burget, M. Karafiát, and J. Černocký, "Phoneme based acoustics keyword spotting in informal continuous speech," *LNAI*, vol. 3658, no. 2005, 2005.

[12] I. Szöke, *Hybrid word-subword spoken term detection*, Ph.D. thesis, Faculty of Information Technology BUT, 2010.

[13] J. Tejedor, I. Szöke, and M. Fapšo, "Novel methods for query selection and query combination in query-by-example spoken term detection," in *Proc. SSCS*, Florence, Italy, 2010.

[14] A. Muscariello, G. Gravier, and F. Bimbot, "A zero-resource system for audio-only spoken term detection using a combination of pattern matching techniques," in *Proc. INTERSPEECH*, Florence; Italy, Aug. 2011, ISCA.

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

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

[17] *MediaEval 2011 Workshop*, Pisa, Italy, Sept. 2011.