HAL
open science

The Spoken Web Search Task

Florian Metze, Etienne Barnard, Marelie Davel, Charl van Heerden, Xavier Anguera, Guillaume Gravier, Nitendra Rajput

To cite this version:

Florian Metze, Etienne Barnard, Marelie Davel, Charl van Heerden, Xavier Anguera, et al.. The Spoken Web Search Task. Working Notes Proceedings of the MediaEval 2012 Workshop, 2012, Italy. hal-00757594

HAL Id: hal-00757594

https://hal.science/hal-00757594v1

Submitted on 27 Nov 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The Spoken Web Search Task

Florian Metze[*] Etienne Barnard[†] Marelie Davel[†] Charl van Heerden[†] Xavier Anguera[‡] Guillaume Gravier[§] Nitendra Rajput[**]

ABSTRACT
In this paper, we describe the "Spoken Web Search" Task, which is being held as part of the 2012 MediaEval campaign. The purpose of this task is to perform audio search in multiple languages, with very little resources being available for each individual language. The data is being taken from audio content that was created in live and realistic low-resource settings.

Categories and Subject Descriptors
H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing.

General Terms
Algorithms, Performance, Experimentation, Languages.

Keywords
Spoken Term Detection, Zero-Resource Techniques.

1. INTRODUCTION
The "Spoken Web Search" task of MediaEval 2012 [3] involves searching *for* audio content *within* audio content *using* an audio content query. The task requires researchers to build a language-independent audio search system so that, given an audio query, it should be able to find the appropriate audio file(s) and the (approximate) location of a query term within the audio file(s). Evaluation is performed using standard NIST metrics [1]. The 2012 evaluation expands on the MediaEval 2011 "Spoken Web Search" task [6], adding more evaluation conditions, and a new "African" dataset as the primary evaluation corpus.

As contrastive condition, participants can submit systems not based on an audio query. Note that language labels and pronunciation dictionaries were not provided for any evaluation data. The lexical form of the query cannot be used to deduce the language in the audio-only condition. The goal of the task is primarily to compare the performance and limitations of different approaches on this type of task and data, not so much a performance comparison between different sites.

2. MOTIVATION
Imagine you want to build a simple speech recognition system, or at least a spoken term detection (STD) or keyword search (KWS) system in a new dialect or language, for which only very few audio examples are available. Maybe there even is no written form for that dialect or language? Is it possible to do something useful (i.e. identify the topic of a query) by using only those very limited resources available? Full speech recognition may be unrealistic to be used for such a task, but it may not be required to solve a specific information access or search problem.

Copyright is held by the author/ owner(s).

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy.

3. RELATED WORK
This task has originally been suggested by IBM Research India, and the 2011 data was provided by this group, see [2]. In 2012, the evaluation is performed on new data [5]; participants are given access to the 2011 data to allow them to gauge their systems.

4. TASK DESCRIPTION
Participants receive development and evaluation utterances (audio data) as well as development and evaluation (audio) queries, described in more detail below. Only the occurrence of development queries in development utterances is provided.

Participants are required to submit the following runs in the audio-only condition, without exploiting the textual form of the queries:

- On the evaluation utterances: identify which query (from the set of development queries) occurs in each utterance (*0-n* matches per term, i.e. not every term necessarily occurs, but multiple matches are possible per utterance)
- On the evaluation utterances: identify which (evaluation) query occurs in each utterance (*0-n* matches)
- On the development utterances: identify which evaluation query occurs in each utterance (*0-n* matches)

The purpose of requiring these three conditions is to see how critical tuning is for the different approaches, i.e., we assume that participants already know their performance for "development queries" on "development utterances", so for evaluation we will evaluate the performance of unseen "evaluation queries" on previously known "evaluation utterances" (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. There may be partial overlap between evaluation and development queries.

Optionally, participants can submit the same runs also using the provided lexical form (not the phonetic transcription) of the query, i.e. they could use existing speech recognition systems, etc., for comparison purposes. Participants can submit multiple systems, but need to designate one primary system. Participants are encouraged to submit a system trained only on data released for the 2012 SWS task, but are allowed to use any additional resources they might have available (with one exception for 2012), as long as their use is documented.

4.1 Development Data
Participants are provided with two distinct datasets. The 2011 "Indian" data has been kindly made available by the Spoken Web team at IBM Research, India [4]. The audio content is spontaneous speech that has been created over phone 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. The data set comprises audio from four different Indian languages: English, Hindi, Gujarati and Telugu. Each data item is ca. 4-30

seconds in length. The development set contains 400 utterances (100 per language) and 64 queries (16 per language), all as audio files recorded in 8kHz/ 16bit, as WAV file. For each query and utterance, the organizers also provided the lexical transcription. For each utterance, the organizers provide 0-n matching queries (but not the exact location of the match).

The 2012 "African" data consists of 1580 audio files (395 per language) from isiNdebele, Siswati, Tshivenda, and Xitsonga, and 100 example queries in these languages with overall similar characteristics to the "Indian" data. These were also collected over a telephone channel [5] and provided as 8kHz/ 16bit WAV files. No language labels or pronunciation dictionaries are provided by default, although these are available for contrastive experiments. The locations of the occurrences are being provided.

4.2 Evaluation Data

For the 2011 "Indian" data, the test set consists of 200 utterances (50 per language) and 36 queries (9 per language) as audio files, with the same characteristics. As with the development data, the lexical form of the query was provided, but not the matching utterances. The 2012 "African" dataset consists of 1660 audio files and 100 queries (transcribed with 170 words), from 4 languages (25 queries per language).

Data is being provided as a "termlist" XML file, in which the "termid" corresponds to the filename of the audio query. This information is packaged together with the scoring software ,e.g.:

```
<?xml version="1.0" encoding="UTF-8"?>

<termlist ecf_filename="expt_06_std_dryrun06_
eng_all_spch_expt_1.ecf.xml" ... >

<term termid="DryRun06_eng_0001"><termtext>
"years"</termtext></term>

</termlist>
```

5. EVALUATION OF RESULTS

The ground truth was created manually by native speakers, and provided by the task organizers, following the principles and using the tools of NIST's Spoken Term Detection (STD) evaluations. The primary evaluation metric was ATWV (Actual Term Weighted Value), as used in the NIST 2006 Spoken Term Detection (STD) evaluation [1].

On development data, the systems can be scored using the software provided in the `me2012-scoring-beta3.tar.bz2` archive available on the FTP server, as described in the MediaEval Task Wiki. This software allows participants to verify themselves that the organizers can process their output for scoring, and reports the respective figure of merit plus graphs on the development data. For evaluation, participants will mail their output on development and evaluation data to the organizers, who compute the results. Because the 2011 evaluation data has already been distributed to the community in the past, it is designated as a contrastive dataset for 2012, and is provided so that participants can compare their results to published work. Participants should however focus on the larger "African" data, and treat the "Indian" data as a "progress" set, even if this use cannot be mandated by the organizers.

For 2012, the NIST-compatible reference files are generated on automatically aligned transcriptions. The trade-off parameters are set so that equal weight is being put on missed detections and false alarms, given the known number of occurrences of search terms in the respective data sets. These parameters are different for each evaluated reference and query set and are set in evaluation scripts. Other weighting schemes may be investigated during the evaluation as diagnostic runs.

6. OUTLOOK

Spoken Web Search and similar technology is primarily targeted at communities whose members do not have Internet access, have low literacy levels, or speak in traditional languages for which good speech technology does not exist. Low (or even zero) resource speech recognition is currently receiving a lot of attention, because it could be a useful contribution in these, and other settings. We will discuss evaluation outcome and future directions at the workshop and in future publications.

7. ACKNOWLEDGMENTS

The organizers would like to thank Martha Larson for organizing this event [3], and the participants for putting in a lot of hard work into submitting their systems. The "African" data [5] has kindly been collected by CSIR and made available by NWU.

8. REFERENCES

[1] J. Fiscus, J. Ajot, J. Garofolo, and G. Doddington, 2007, "*Results of the 2006 Spoken Term Detection Evaluation*," Proc. ACM SIGIR 2007, Workshop in Searching Spontaneous Conversational Speech (SSCS).

[2] M. Diao, S. Mukherjea, N. Rajput, and K. Srivastava, *"Faceted Search and Browsing of Audio Content on Spoken Web,"* Proc. CIKM 2010.

[3] http://www.multimediaeval.org/mediaeval2012/index.html

[4] http://domino.research.ibm.com/comm/research_projects.nsf/pages/pyrmeait.index.html

[5] E. Barnard, M. Davel, and C. van Heerden, "ASR Corpus design for resource-scarce languages," in Proc. INTERSPEECH, Brighton, UK; Sep. 2009, pp. 2847-2850.

[6] F. Metze, N. Rajput, X. Anguera, M. Davel, G. Gravier, C. v. 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; Mar. 2012. IEEE.

* Carnegie Mellon University; Pittsburgh, PA, U.S.A; `fmetze@cs.cmu.edu`

† North-West University; Vanderbijlpark, South Africa; `{etienne.barnard|marelie.davel|cvheerden}gmail.com`

‡ Telefonica Research; Barcelona, Spain; `xanguera@tid.es`

§ IRISA; Rennes, France; `guig@irisa.fr`

** IBM Research; New Delhi, India; `rnitendra@in.ibm.com`