The Spoken Web Search Task

Xavier Anguera[*] Florian Metze[†] Andi Buzo[‡] Igor Szoke[#] Luis Javier Rodriguez-Fuentes[§]

ABSTRACT

In this paper, we describe the "Spoken Web Search" Task, which is being held as part of the 2013 MediaEval campaign. The purpose of this task is to perform audio search in multiple languages and acoustic conditions, with very few resources being available for each individual language. This year the data contains audio from nine different languages and is much bigger in size than in previous years, mimicking realistic low/zero-resource settings.

1. INTRODUCTION

The "Spoken Web Search" (SWS) task of MediaEval 2013 [3] involves searching *for* audio content *within* audio content *using* an audio 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 exact location(s) of a query term within these audio file(s). Evaluation is performed using standard NIST metrics [1] in addition to some other indicators.

The 2013 evaluation expands on the MediaEval 2011 and 2012 "Spoken Web Search" tasks [6,7] by increasing the size of the test dataset and the number of languages (which were recorded in different acoustic conditions). In addition, a baseline system is being offered this year to first-time participants as a virtual kitchen appliance.

2. MOTIVATION and RELATED WORK

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, language or acoustic condition, for which only very few audio examples are available. Maybe there even are no transcripts available for that data. Is it possible to do something useful (e.g. identify the topic of a query) by using only those very limited resources available? Full-fledged speech recognition may be unrealistic to be used for such a task, which may not be required to solve a specific information access or search problem.

This task was originally proposed by IBM Research India, who provided the 2011 data [2]. In 2012, the evaluation was performed on new data gathered from 4 different African languages [5]. The 2012 data is made available to participants to help them in their system development.

3. TASK DESCRIPTION

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

Participants are required to identify and submit which query (or queries, from the set of evaluation queries) occur(s) in each utterance (*0-n* matches per term, i.e. not every term necessarily occurs, but multiple matches are possible per utterance). There may be partial overlap between evaluation and development queries. In addition, participants are asked to submit their development output (i.e. the detection of development queries on the data) 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 2013 SWS task, but are allowed to use any additional resources they might have available, as long as their use is documented.

For the first time this year, a "Speech Recognition Virtual Kitchen" appliance [8] is made available to participants as a baseline system to experiment with. This consists of a Linux-based virtual machine, running a complete SWS system.

3.1 Development and evaluation Data

As a result of a joint effort between several institutions, a challenging new dataset, together with accompanying queries, has been put together for the 2013 evaluation. This dataset is composed of 20 hours of audio in the following 9 languages: Albanian, Basque, Czech, non-native English, Isixhosa, Isizulu, Romanian, Sepedi and Setswana. The recording acoustic conditions are not constant for all languages; some being obtained from in-room microphone recordings while others have been obtained through street recordings with cellphones. All data has been converted to 8KHz/ 16bit WAV files. Moreover, the amount of audio available for each language is not the same for all languages. Such database is over 5 times the size of the 2012 databases. The development and evaluation queries are mutually exclusive segments defined within the same data collection. For this reason, no information on the language being spoken or the transcription of the files is released with the development runs. We believe that with such a variety of data the concept of over-fitting to the dev-test set is quite diluted and, if any, it should be seen as a good thing for systems to be able to take advantage from knowing the possible acoustics of the test languages.

Accompanying the dataset, two sets of queries have been created for use in the development and evaluation, each one containing two subsets of basic and extended queries. A basic set of 500+ queries each are to be used by participants in their required runs. In addition, for some of the basic queries, alternative spoken instances of the same lexical terms have also been gathered and are made available to participants to be used (together with the basic queries) in their extended runs. Such extended runs are intended to represent how results would vary if systems could take advantage of multiple repeated queries.

In addition to the main database used for this year, the 2012 "African" database [4] is also being made available to participants in hope it is of help in the development phase. It consists of over 1580 files and 100 queries both for development and evaluation, recorded in 4 African Languages. Participants should note that the acoustic conditions of this dataset only match those of a small part of the 2013 dataset.

A "termlist" XML file and a transcription RTTM file are provided with the development data, following the guidelines of the NIST-STD 2006 evaluation [1]. For this year the reference files do not contain any information regarding the language or the content spoken in each file, and only the locations of the queries is given

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain.

in the reference RTTM file. This is done so in order not to give away any extra information about the dataset when releasing the development data, as it is shared with the evaluation queries.

4. EVALUATION OF RESULTS

The ground truth for this year has been created in a variety of ways. Sometimes is has been created manually by native speakers while in other cases a speech recognition system has been used to force-align the transcripts at word level. Note that word alignments might not be perfect, which is why a margin of error is allowed by the scoring scripts.

The main evaluation metric this year remains the same as previous years by following the principles and using the tools of NIST's Spoken Term Detection (STD) evaluations. The primary evaluation metric is ATWV (Actual Term Weighted Value), as used in the NIST 2006 Spoken Term Detection (STD) evaluation [1]. A scoring package with easy-to-use scripts and an example scoring setup have been made available to participants with the development data. This year we are again applying a different scoring working point by modifying the miss and false alarm costs to better match the new test data.

In addition, two secondary metrics are being introduced this year. On the one hand, the normalized cross-entropy metric C_{nxe} evaluates the information provided by system scores (in contrast to TWV, which uses system decisions). This metric originates from the NIST SRE evaluations and is computed assuming that submitted scores can be interpreted as log-likelihood ratios. On the other hand, the real-time factor evaluates the required resources used by the systems. In addition, participants are requested to indicate the type of machines used in the evaluation and (approximately) the peak memory usage in order for organizers to compute a global processing load metric per system. See [9] for a detailed description on these metrics.

5. OUTLOOK

Low (or even zero) resource speech recognition is currently receiving a lot of attention and will soon reach maturity to be useful for real-life scenarios. The "Spoken Web Search" task originated as an alternative to standard techniques for low/zero-resourced languages where good speech recognizers do not exist. This year we have extended this paradigm to include audio data for which not much is known a priori, by mixing several languages and acoustic conditions in the same test dataset. By comparing the results obtained by the different systems in this friendly evaluation we expect to help push forward the state-of-the-art in this area.

6. ACKNOWLEDGMENTS

The SWS task organizers would like to thank the Mediaeval organizers [3] and all the participants for putting in a lot of hard work into submitting their systems. The "African" data [5] for this year and last year has kindly been collected by CSIR and made available by Charl van Heerden at NWU. Igor Szöke was supported by Grant Agency of Czech Republic post-doctoral project No.GPP202/12/P567

7. 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/mediaeval2013/index.html

[4] F. Metze, E. Barnard, M. Davel, C.V.Heerden, X.Anguera G. Gravier and N. Rajput, "The spoken web search task", in Workshop notes of Mediaeval 2012, Pisa, Italy

[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.

[7] F. Metze, X. Anguera, E. Barnard, M. Davel and G. Gravier. The Spoken Web Search task at MediaEval 2012. In Proc. ICASSP, Vancouver; May. 2013. IEEE.

[8] F. Metze and E. Fosler-Lussier, "The Speech Recognition Virtual Kitchen: An Initial Prototype", in Proc. Interspeech 2012, Portland, USA

[9] Luis J. Rodriguez-Fuentes, M. Penagarikano, "MediaEval 2013 Spoken Web Search Task: System Performance Measures", n.TR-2013-1, Department of Electricity and Electronics, University of the Basque Country, 2013. Link: http://gtts.ehu.es/gtts/NT/fulltext/rodriguezmediaeval13.pdf

* Telefonica Research; Barcelona, Spain; xanguera@tid.es

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

‡ University Politehnica of Bucharest, Romania, andi.buzo@upb.ro

♯ Brno University of Technology, Czech Republic; szoke@fit.vutbr.cz

§ University of the Basque Country, Spain; luisjavier.rodriguez@ehu.es