Query by Example Search on Speech at Mediaeval 2014

Xavier Anguera
Telefonica Research
Barcelona , Spain
xanguera@tid.es

Luis Javier Rodriguez-Fuentes
University of the Basque Country
Leioa, Spain
luisjavier.rodriguez@ehu.es

Igor Szöke
Brno University of Technology
Brno, Czech Republic
szoke@fit.vutbr.cz

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

Florian Metze
Carnegie Mellon University
Pittsburgh, PA, U.S.A
fmetze@cs.cmu.edu

ABSTRACT

In this paper, we describe the "Query by Example Search on Speech Task" (QUESST, formerly SWS, "Spoken Web Search"), held as part of the MediaEval 2014 evaluation campaign. As in previous years, the proposed task requires performing language-independent audio search in a low resource scenario. This year, the task has been designed to get as close as possible to a practical use case scenario, in which a user would like to retrieve, using speech, utterances containing a given word or short sentence, including those with limited inflectional variations of words, some filler content and/or word re-orderings.

1. INTRODUCTION

After three years running as SWS ("Spoken Web Search") [4, 3, 1, 2], the task has been renamed to QUESST ("QUery by Example Search on Speech Task") to better reflect its nature: to search FOR audio content WITHIN audio content USING an audio query. As in previous years, the search database was collected from heterogeneous sources, covering multiple languages, and under diverse acoustic conditions. Some of these languages are resource-limited, some are recorded in challenging acoustic conditions and some contain heavily accented speech (typically from non-native speakers). No transcriptions, language tags or any other metadata are provided to participants. The task therefore requires researchers to build a language-independent audio-to-audio search system. As in previous years, the database will be made publicly available for research purposes after the evaluation concludes.

Three main changes were introduced for this year's evaluation, namely on the the *search task*, on the *evaluation metrics*, and on the *types of query matchings*. First, the task no longer requires the localization (time stamps) of query matchings within audio files (which, on the other hand, are relatively short: less than 30 seconds long). However, systems must provide a score (a real number) for each query matching, the higher (the more positive) the score, the more likely that the query appears in the audio file. Second, the *normalized cross entropy cost* (C_{nxe}) [5] is used as the primary metric, whereas the Actual Term Weighted Value (ATWV, used as primary metric in previous years) is kept as a secondary metric for diagnostic purposes, which means that systems must provide not only scores, but also Yes/No decisions. And third, three types of query matchings are considered: the first one is the "exact match" case used in previous years, whereas the second one, which allows for inflectional variations of words, and the third one, which allows for word re-orderings and some filler content between words, are "approximate matches" that simulate how we imagine that users would want to use this technology.

2. BRIEF TASK DESCRIPTION

QUESST is part of the Mediaeval 2014 evaluation campaign[1]. As usual, two separate sets of queries are provided, for development and evaluation, along with a single set of audio files, on which both sets of queries must be searched on. The set of development queries and the set of audio files are distributed early (June 2nd), including the groundtruth and the scoring scripts, for the participants to develop and evaluate their systems. The set of evaluation queries is distributed one month later (July 1st). System results (*for both sets of queries*) must be returned by the evaluation deadline (September 9th), including a likelihood score and a Yes/No decision for each pair (query, audio file). Note that not every query necessarily appears in the set of audio files, and that several queries may appear in the same audio file. Also, there could be some overlap between evaluation and development queries. Multiple system results can be submitted (up to 5), but one of them (presumably the best one) must be identified as *primary*. Also, although participants are encouraged to train their systems using only the data released for this year's evaluation, they are allowed to use any additional resources they might have available, as long as their use is documented in their system papers. System results are then scored and returned to participants (by September 16th), who must prepare a working notes (two-page) paper describing their systems and return it to the organizers (by September 28th). Finally, systems are presented and results discussed in the Mediaeval workshop, which serves to meet fellow participants, to share ideas and to bootstrap future collaborations.

[1] http://www.multimediaeval.org/mediaeval2014/

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

3. THE QUESST 2014 DATASET

The QUESST 2014 dataset[2] is the result of a joint effort by several institutions to put together a sizable amount of data to be used in this evaluation and for later research on the topic of query-by-example search on speech. The search corpus is composed of around 23 hours of audio (12492 files) in the following 6 languages: Albanian, Basque, Czech, non-native English, Romanian and Slovak, with different amounts of audio per language. The search utterances, which are relatively short (6.6 seconds long on average), were automatically extracted from longer recordings and manually checked to avoid very short or very long utterances. The QUESST 2014 dataset includes 560 development queries and 555 evaluation queries, the number of queries per language being more or less balanced with the amount of audio available in the search corpus. A big effort has been made to manually record most of the queries, in order to avoid problems observed in previous years due to acoustic context derived from cutting the queries from longer sentences. Speakers recruited for recording the queries were asked to maintain a normal speaking speed and a clear speaking style. All audio files are PCM encoded at 8 kHz, 16 bits/sample, and stored in WAV format.

4. THE GROUND-TRUTH

The biggest novelty in this year's evaluation comes from the new (relaxed) concept of a query match, which strongly affects the ground-truth definition and thus the way systems are expected to work. Besides the "exact matching" used in previous years, two types of "approximate matchings" are considered. We denote these matchings as of Type 1, 2 and 3, respectively, and are defined as follows:

Type 1 (Exact): Only occurrences that exactly match the lexical representation of the query are considered as hits, just like in previous years. For example, the query "white horse" would match the utterance "My white horse is beautiful".

Type 2 (Variant): In this case, query occurrences that slightly differ from its lexical representation, either at the beginning or at the end of the query, are considered as hits. Systems therefore need to account for small portions of audio that do not match its lexical representation. When producing the ground-truth for this type of matchings, the matching part of any query was required to exceed 5 phonemes (250 ms), and the non-matching part was required to be much smaller than the matching part. For example, the query "researcher" would match an audio file containing "research" (note that the query "research" would also match an audio file containing "researcher").

Type 3 (Reordering/Filler): Given a multi-word query, a hit is required to contain all the words in the query, but possibly in a different order and with some small amount of *filler content* between words; slight differences between word occurrences and their lexical representations are also allowed (like in Type 2). For example the query "white snow" would match an utterance containing either "snow is white", "whitest snow" or "whiter than snow". Note that queries provided in this evaluation are spoken continuously, with no silences between words, and thus participants should develop robust techniques to account for partial matchings. Note also that, when producing the ground-truth for this type of matchings, hits are were allowed to contain a large amount of filler content between words.

The ground truth was created either manually by native speakers or automatically by speech recognition engines tuned to each particular language, and provided by the task organizers, following the format of NIST's Spoken Term Detection evaluations. The development package contains a general ground-truth folder (the one that must be used to score system results on the development set) which considers all types of matchings, but also three ground-truth folders specific to each type of matchings, to allow participants evaluate their progress on each condition during system development.

5. PERFORMANCE METRICS

The primary metric used in QUESST 2014 is the *normalized cross entropy cost* (C_{nxe}), already used in SWS 2013 as a secondary metric [1]. This metric has been used for several years in the language and speaker recognition fields to calibrate system scores, and shows interesting properties. Furthermore, we found experimentally that C_{nxe} and ATWV performances correlate quite well. A scoring script has been specifically prepared for this year's evaluation, so that NIST software is not required anymore[3]. For the C_{nxe} scores to be meaningful, participants are requested either to return a score (that will be taken as a log-likelihood ratio) for every pair (query, audio file), or alternatively, to define a default (floor) score for all the pairs not included in the results file. TWV metrics are computed with the following application parameters: $P_{target} = 0.0008$, $C_{fa} = 1$ and $C_{miss} = 100$. Participants are also required to report on their real-time running factor, hardware characteristics and peak memory requirements, in order to profile the different approaches applied. See [5] for further information on how the metrics work and how they are computed.

6. ACKNOWLEDGEMENTS

We would like to thank the Mediaeval organizers for their support and all the participants for their hard work. Data were provided by QUESST organizers and by the Technical University of Kosice (TUKE), Slovak Republic. Igor Szöke was supported by the Czech Science Foundation, under project GPP202/12/P567.

7. REFERENCES

[1] X. Anguera, F. Metze, A. Buzo, I. Szoke, and L. J. Rodriguez-Fuentes. The Spoken Web Search Task. In *Proc. Mediaeval 2013 Workshop*, 2013.

[2] F. Metze, X. Anguera, E. Barnard, M. Davel, and G. Gravier. Language Independent Search in MediaEval's Spoken Web Search Task. *Computer Speech and Language, Special Issue on Information Extraction & Retrieval*, 2014.

[3] F. Metze, E. Barnard, M. Davel, C. van Heerden, X. Anguera, G. Gravier, and N. Rajput. The Spoken Web Search Task. In *Proc. Mediaeval 2012 Workshop*, 2012.

[4] N. Rajput and F. Metze. Spoken Web Search. In *Proc. Mediaeval 2011 Workshop*, 2011.

[5] L. J. Rodriguez-Fuentes and M. Penagarikano. MediaEval 2013 Spoken Web Search Task: System Performance Measures. Technical Report TR-2013-1, DEE, University of the Basque Country, 2013. Online: http://gtts.ehu.es/gtts/NT/fulltext/rodriguezmediaeval13.pdf.

[2] A download link will be provided after the evaluation.

[3] Thanks to Mikel Peñagarikano, from the University of the Basque Country, for creating the scoring script.