Query by Example Search on Speech at Mediaeval 2015

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

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

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

Xavier Anguera
Sinkronigo S.L.
Barcelona, Spain
xanguera@gmail.com

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

Jorge Proenca
University of Coimbra
Coimbra, Portugal
jproenca@co.it.pt

Martin Lojka
Technical University of Kosice
Kosice, Slovakia
martin.lojka@tuke.sk

Xiao Xiong
Temasek laboratories, Nanyang Technological University
Singapore
xiaoxiong@ntu.edu.sg

ABSTRACT

In this paper, we describe the "Query by Example Search on Speech Task" (QUESST), held as part of the MediaEval 2015 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. We also stressed a mismatch caused by noise and reverberations.

1. INTRODUCTION

This is the fifth year of query-by-example search on speech evaluations [9, 6, 1, 2]. The task of QUESST ("QUery by Example Search on Speech Task") is 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.

Compared to the previous year, two main changes were introduced for this year's evaluation. First, we provide queries with longer context. So participants can use this surrounding speech to adapt their systems. Second, we artificially add noises and reverberations into the data. This aims to measure robustness of particular feature extractions and algorithms in heavy channel mismatch.

As in the previous year, the proposed task does not require the localization (time stamps) of query matchings within audio files. However, systems must provide a score (a real number) for each query matching. The higher (the more positive) the score, the more likely it is that the query appears in the audio file. The *normalized cross entropy cost* (C_{nxe}) [3, 10] is used as the primary metric, whereas the Actual Term Weighted Value (ATWV) is kept as a secondary metric for diagnostic purposes, which means that systems must provide not only scores, but also Yes/No decisions. Three types of query matchings are considered: the first one (*T1*) involves "exact matches" whereas the second one (*T2*) allows for inflectional variations of words or word re-orderings (that is, "approximate matches"); the third one (*T3*) is similar to *T2*, but queries are drawn from conversational speech, thus containing strong coarticulations and some filler content between words.

2. BRIEF TASK DESCRIPTION

QUESST is part of the Mediaeval 2015 evaluation campaign[1]. As usual, two separate sets of queries were provided, for development and evaluation, along with a single set of audio files, on which both sets of queries had to be searched on. The set of development queries and the set of audio files were distributed early (April 1st), including the groundtruth and the scoring scripts, for the participants to develop and evaluate their systems. The set of evaluation queries was distributed one month later (May 1st). System results (*for both sets of queries*) had to be returned by the evaluation deadline (July 22nd), 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.

Multiple system results could be submitted (up to 5), but one of them (presumably the best one) had to be identified as *primary*. Also, although participants were encouraged to train their systems using only the data released for this year's evaluation, they were allowed to use any additional resources they might have available, as long as their use was documented in their system description papers. System results were then scored and returned to participants (by July 29th), who had to prepare a working notes (two-page) paper describing their systems and return it to the organizers (by August 28th). Finally, systems were presented and results

[1] http://www.multimediaeval.org/mediaeval2015/

MediaEval 2015 Workshop, September 14-15, 2015, Wurzen, Germany.

discussed in the Mediaeval workshop, which serves to meet fellow participants, to share ideas and to bootstrap future collaborations.

3. THE QUESST 2015 DATASET

The QUESST 2015 dataset 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 18 hours of audio (11662 files) in the following 7 languages: Albanian, Czech, English, Mandarin, Portuguese, Romanian and Slovak [8], with different amounts of audio per language. The search utterances, which are relatively short (5.8 seconds long on average), were automatically extracted from longer recordings and manually checked to avoid very short or very long utterances. The QUESST 2015 dataset includes 445 development queries and 447 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 the queries, in order to avoid problems observed in previous years due to acoustic context derived from cutting 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.

The data was then artificially noised and reverberated with equal amounts of clean, noisy, reverb and noisy+reverb speech. We used both stationary and transient noises downloaded from `https://www.freesound.org`. Reverberation was obtained by passing the audio through a filter with an artificially generated room impulse response (RIR) [5].

4. THE GROUND-TRUTH

Similarly to the last year's evaluation, we have applied a relaxed concept of a query match, which strongly affects the ground-truth definition and thus the way systems are expected to work. Besides "exact matchings" (Type 1), two types of "approximate matchings" (Types 2 and 3) are considered, which are defined as follows:

Type 1 (Exact match): Occurrences of single or multiple word queries in utterances should exactly match the lexical representation of the query. An example of this case is the query "white horse" that should match the utterance "My white horse is beautiful" but should not match to "The whiter horse is faster".

Type 2 (Re-ordering and small lexical variations): Here the search algorithm should cope with:

- Lexical variations. Occurrences of single/multiple word queries might differ slightly, with small portions of audio either at the beginning or the end of the segment that do not match the lexical form of the reference. An example of this type of search would be "researcher" matching an utterance containing "research" (note that the inverse would also be possible).
- Word re-orderings and small filler content between words. For example, when searching for the query "white horse", systems should be able to match both "My horse is white" and "I have two white and beautiful horses". Note that the matching words may also contain slight variations with regard to the lexical form of the query.

Type 3 (Conversational queries in context): This type of search is another step towards realistic use-case scenarios. In this case, the spoken query is just part of a sentence, that may contain silent/filled pauses and irrelevant words. For example, "Google, let me find some red [uh] white [pau] horse to ride today" could be one of these complex queries. As it is extremely difficult to distinguish between query words ("white [pau] horse") and "fillers" ("Google, let me find some red [uh]" and "to ride today"), we provide timing meta-data of the relevant segment inside the spoken query.

As Type 3 queries required timing meta-data, all the queries were recorded within a context and timing information was provided for all of them. The context of Types 1 and 2 was "artificial": speakers were asked to say several words before and after the query, with significant pauses around the query to avoid coarticulations. Participants are free to use the "context" of the spoken query (e.g. for adaptation).

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

In QUESST 2015, C_{nxe} and ATWV are used as primary and secondary metrics, respectively. 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. 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 [10] for further information on how the metrics work and how they are computed.

6. PROVIDED TOOLS

We offered some of the basic tools for paricipants to make their "first contact" with the QUESST easier. We provided Bottle-Neck feature [4] extraction tool trained on Russian and Hungarian Speechdat-E. Next, calibration and fussion script based on logistic regresion [11] and DTW search [12], both developed at BUT, were provided. Finally, the data and all the scripts were setup in a Virtual Machine which was provided to the participants through the Speech Recognition Virtual Kitchen (`http://speechkitchen.org/`, [7]).

7. ACKNOWLEDGEMENTS

We would like to thank the Mediaeval organizers for their support and all the participants for their hard work. Portuguesse data were provided with thanks to Tecnovoz project PMDT No. 03/165

8. 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*.

[2] X. Anguera, J. L. Rodriguez-Fuentes, I. Szoke, A. Buzo, and F. Metze. Query by Example Search on Speech at Mediaeval 2014. In *Proc. Mediaeval 2014 Workshop*.

[3] X. Anguera, L.-J. Rodriguez-Fuentes, A. Buzo, F. Metze, I. Szoke, and M. Penagarikano. QUESST2014: Evaluating Query-by-Example Speech Search in a Zero-Resource Setting with Real-Life Queries. In *Proc. ICASSP*, pages 5833–5837, 2015.

[4] F. Grézl, M. Karafiát, S. Kontár, and J. Černocký. Probabilistic and bottle-neck features for lvcsr of meetings. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, pages 757–760. IEEE Signal Processing Society, 2007.

[5] S. G. McGovern. Fast image method for impulse response calculations of box-shaped rooms. *Applied Acoustics*, 70(1):182–189, 2009.

[6] 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*.

[7] F. Metze, E. Riebling, E. Fosler-Lussier, A. Plummer, and R. Bates. The speech recognition virtual kitchen turns one. In *to appear in Proc. INTERSPEECH*. ISCA, 2015.

[8] M. Pleva and J. Juhar. Tuke-bnews-sk: Slovak broadcast news corpus construction and evaluation. In N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

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

[10] 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.

[11] I. Szoke, L. Burget, F. Grézl, J. Černocký, and L. Ondel. Calibration and fusion of query-by-example systems - but sws 2013. In *Proceedings of ICASSP 2014*, pages 7899–7903. IEEE Signal Processing Society, 2014.

[12] I. Szoke, M. Skacel, L. Burget, and J. H. Cernocky. Coping with Channel Mismatch in Query-by-Example - BUT QUESST 2014. In *Proc. ICASSP*, pages 5838–5843, 2015.