Information Processing and Management 56 (2019) 43–60

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Efficient query-by-example spoken document retrieval combining phone multigram representation and dynamic time warping

Paula Lopez-Otero*, Javier Parapar, Alvaro Barreiro

Universidade da Coruña - CITIC, Facultad de Informática Campus de Elviña S/N, A Coruña 15071, Spain

ARTICLE INFO

Keywords:
Query-by-example spoken document retrieval
Phone decoding
Phone n-grams
Phone posteriorgrams
Dynamic time warping

ABSTRACT

Query-by-example spoken document retrieval (QbESDR) aims at finding those documents in a set that include a given spoken query. Current approaches are, in general, not valid for real-world applications, since they are mostly focused on being effective (i.e. reliably detecting in which documents the query is present) but practical implementations must also be efficient (i.e. the search must be performed in a limited time) in order to allow for a satisfactory user experience. In addition, systems usually search for exact matches of the query, which limits the number of relevant documents retrieved by the search. This paper proposes a representation of the documents and queries for QbESDR based on combining different-sized phone n-grams obtained from automatic transcriptions, namely phone multigram representation. Since phone transcriptions usually have errors, several hypotheses for the query transcriptions are combined in order to ease the impact of these errors. The proposed system stores the document in inverted indices, which leads to fast and efficient search. Different combinations of the phone multigram strategy with a state-of-art system based on pattern matching using dynamic time warping (DTW) are proposed: one consists in a two-stage system that intends to be as effective but more efficient than a DTW-based system, while the other aims at improving the performance achieved by these two systems by combining their output scores. Experiments performed on the MediaEval 2014 Query-by-Example Search on Speech (QUESST 2014) evaluation framework suggest that the phone multigram representation for QbESDR is a successful approach, and the assessed combinations with a DTW-based strategy lead to more efficient and effective QbESDR systems. In addition, the phone multigram approach succeeded in increasing the detection of non-exact matches of the queries.

1. Introduction

The interaction with spoken contents has increased dramatically in the last few years due to the proliferation of audiovisual documents that are part of our daily life. This new paradigm of communication demands strategies for searching for contents of interest, creating the need for tools that allow the retrieval of spoken documents, task known as spoken document retrieval (SDR) (Sparck Jones, Jones, Foote, & Young, 1996). SDR can be carried out using either written or spoken queries. This latter approach, known as query-by-example SDR (QbESDR), allows the communication with devices in a natural manner while easing the access to such technologies to visually impaired users.

The approaches for QbESDR found in the literature can be divided into two main groups: those based on automatic speech recognition (ASR), which imply transcribing both documents and queries into words or sub-words (van Heerden, Karakos,

* Corresponding author.
E-mail addresses: paula.lopez.otero@udc.gal (P. Lopez-Otero), javierparapar@udc.gal (J. Parapar), barreiro@udc.gal (A. Barreiro).

https://doi.org/10.1016/j.ipm.2018.09.002
Received 20 February 2018; Received in revised form 31 July 2018; Accepted 7 September 2018

Narasimhan, Davel, & Schwartz, 2017; Martinez et al., 2014; Nakagawa, Iwami, Fujii, & Yamamoto, 2013; Sakamoto, Yamamoto, & Nakagawa, 2014; Xu et al., 2016; 2015); and those that make use of pattern matching techniques, usually by finding alignments of the queries in the documents using the dynamic time warping (DTW) algorithm (Sakoe & Chiba, 1978) or any of its variants (Anguera, 2013; Anguera & Ferrarons, 2013; Mantena, Achanta, & Prahallad, 2014; Müller, 2007). The main limitation of ASR-based strategies is the need for ASR resources in the language of interest, while pattern matching techniques are usually inefficient in terms of computational cost (Anguera & Ferrarons, 2013). In addition, both strategies for QbESDR share an important limitation: they are intended to search for exact matches of the query. This constraint does not recreate a real-world scenario, since a user might want to search for the exact query but also for lexical variations of it. In addition, when looking for queries with multiple terms, the documents that include the terms in a different order may be relevant for the search as well (for example, “president of Brazil” versus “Brazilian president”).

Research in the field of QbESDR has been recently boosted by the organization of competitive evaluations such as Spoken Web Search (Anguera, Metze, Buzo, Szöke, & Rodriguez-Fuentes, 2013; Metze, Barnard et al., 2012; Metze, Rajput et al., 2012) and Query by Example Search on Speech task (Anguera, Rodriguez-Fuentes, Szöke, Buzo, & Metze, 2014; Szöke et al., 2015) at MediaEval; SpokenQuery&Doc Task at NTCIR (Akiba, Nishizaki, Nanjo, & Jones, 2014; 2016); or query-by-example spoken term detection evaluation at Albayzín campaigns (Tejedor & Toledano, 2016; Tejedor et al., 2013; Tejedor, Toledano, Lopez-Otero, Docio-Fernandez, & Garcia-Mateo, 2016). The zero resource speech challenge (Dunbar et al., 2017; Versteegh et al., 2015) is devoted to unsupervised discovery of subword and word units from raw speech, which has QbESTD as one of its applications. The literature related to these evaluations shows a trend that consists in fusing the scores of the detections of different systems (Hou et al., 2015; Lopez-Otero, Docio-Fernandez, & Garcia-Mateo, 2015a; Proença, Castela, & Perdigão, 2015; Szöke, Burget, Grézl, & Ondel, 2013; Yang et al., 2014), which boosts the performance of the individual systems at the cost of increasing the computational demands of the search procedure. When considering a practical implementation for real-world scenarios, QbESDR approaches must be effective (i.e. they must be able to reliably detect in which documents the query is present) but also efficient (i.e. the search must be performed in a limited time) in order to allow for a satisfactory user experience. Hence, massive fusions can be effective but not efficient in practical terms, so new paradigms for QbESDR must be explored.

Two main contributions are presented in this paper, which aim at obtaining effective and efficient systems for real-world QbESDR applications:

- A novel approach for QbESDR based on phone n-gram representation, namely phone multigram representation, is proposed. Given a set of documents, their transcriptions are stored in inverted indices using different sizes of phone n-grams, i.e. the documents are stored tokenized in 1-grams, 2-grams and so forth. Afterwards, for each query, its equivalent tokenization in phone n-grams of different sizes is obtained in order to look for each term in the appropriate index, producing a score that indicates how likely the given set of phone multigrams is present in each document. Additionally, in order to reduce the impact of transcription errors, several transcription hypotheses per query are obtained and searched, which leads to more reliable scores. This approach has several advantages:
  - A small amount of time is necessary for indexing and searching thanks to the efficiency of the inverted indexing and searching procedures.
  - Using phone multigrams for speech representation makes it possible to avoid taking into account the position where the match of each phone n-gram was found, since the smaller likelihood of matching long n-grams compensates that of matching short n-grams.
  - Since the order of the matching n-grams is not considered, the probability of finding non-exact matches of the queries increases.
  - This strategy can be used in a cross-lingual manner, since the language of the phone decoder used to obtain phone transcriptions does not necessarily have to match the language spoken in the documents and queries.

  This approach is inspired by Harding, Croft, and Weir (1997) and Parapar, Freire, and Barreiro (2009), where a similar strategy was used for text retrieval in noisy documents obtained by optical character recognition (OCR). The application scenario is very similar, since both OCR and phone transcriptions have errors that do not allow the search for exact matches of a query.
- Two different combinations of the phone multigram approach with a strategy based on DTW are presented:
  - The first combination consists in a two-stage system: first, the phone multigram system is used to look for candidate matches of the queries in the documents; then, these matches are re-scored using a DTW-based strategy in order to decide whether to keep them or discard them. This strategy increases the efficiency of the search process, since the number of query-document pairs that have to be evaluated with DTW (which is significantly more costly than the phone multigram strategy) is reduced to a great extent.
  - A second combination is proposed, which consists in fusing the output scores of the phone multigram and DTW-based systems. Given that the phone multigram approach is computationally efficient, running these two systems hardly affects the efficiency of the search. Moreover, this combination leads to a relevant improvement in terms of effectiveness as a result of combining different pieces of evidence produced by heterogeneous systems.

The rest of this paper is organized as follows: Section 2 describes the related work; Section 3 presents the phone multigram approach for QbESDR; Section 4 overviews the DTW-based system used in this work; Section 5 presents two techniques for combining the two aforementioned approaches; Section 6 describes the experimental framework; experimental results and a discussion are presented in Section 7; Section 8 reviews other results reported in the literature for the experimental framework used in Section 7; lastly, conclusions and future work are summarized in Section 9.