Speech Communication 84 (2016) 24–35

Contents lists available at ScienceDirect

Speech Communication

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

Finding relevant features for zero-resource query-by-example search on speech

Paula Lopez-Otero[*], Laura Docio-Fernandez, Carmen Garcia-Mateo

AtlantTIC Research Center, E.E. Telecomunicación Campus Universitario, Vigo S/N 36310, Spain

ARTICLE INFO

Article history:
Available online 18 August 2016

Keywords:
Query-by-example search on speech
Feature selection
Dynamic time warping

ABSTRACT

Zero-resource query-by-example search on speech strategies have raised the interest of the research community, as they do not imply training (and therefore, large amounts of training data) or any knowledge about either the language to be processed or any others. These systems usually rely on Mel-frequency cepstral coefficients (MFCCs) for speech representation and dynamic time warping (DTW) or any of its variants for performing the search. Nevertheless, which features to use in this task is still an open research problem, and the use of large feature sets combined with feature selection approaches have not been addressed yet in the query-by-example search on speech scenario. In this paper, we present two methods to select the most relevant features among a large set of acoustic features, for the purpose of estimating the relevance of each feature using the costs of the best alignment path (obtained when performing DTW) and their neighbouring region. To prove the validity of these methods, experiments were carried out in four different search on speech scenarios that were used in international benchmarks, namely Albayzin 2014 search on speech evaluation, MediaEval spoken web search SWS 2013, and MediaEval query-by-example search on speech QUESST2014 and QUESST2015. Experimental results showed a dramatic improvement in the results when reducing the feature set using the proposed techniques, especially in the case of the relevance-based approaches. A comparison between the proposed methods and other representations such as MFCCs, phonetic posteriorgrams and dimensionality reduction based on principal component analysis, showed that the zero-resource approaches presented in this paper are promising, as they outperformed more extended approaches in all the experimental scenarios. The feature relevance estimation approaches, apart from improving search on speech results, also revealed features other than MFCCs that seemed to be a value-added in query-by-example tasks.

1. Introduction

The amount of available spoken documents is steadily increasing, which leads to the need for tools to perform automatic searches within large audio databases. These tools work in two different manners: a search of written queries, which implies transcribing the spoken documents, either manually or automatically; and a search of spoken queries (i.e. searching for audio content, within audio content, using an audio content query (Metze et al., 2012b)). The former approach is commonly issued by performing automatic speech recognition (ASR) of the documents; however, this approach cannot be considered in situations where an ASR system is not available for a given language, when the amount of annotated training data is not enough to train it (Chelba et al., 2008), or when dealing with multilingual scenarios, to cite some examples. In these situations, an alternative solution consists in producing lower level transcriptions such as phonetic transcriptions; this allows the use of phoneme recognizers trained on a different language and leading to cross-lingual approaches, which makes these technologies applicable in under-resourced languages. The search of spoken queries, known as query-by-example search on speech, is commonly approached using pattern matching techniques. The flexibility of these methods has attracted the interest of the research community, leading to the organization of international competitions, in which two different modalities can be distinguished: query-by-example spoken term detection (QbE STD), where the result of the search is the time instants of the occurrences of the query in the search documents (Anguera et al., 2013; Metze et al., 2012a; 2012b; Tejedor et al., 2013; 2014); and query-by-example spoken document retrieval (QbE SDR), where the result of the search is a list of those documents were the query was found (Anguera et al., 2014a; Szöke et al., 2015).

[*] Corresponding author.
E-mail address: plopez@gts.uvigo.es (P. Lopez-Otero).

http://dx.doi.org/10.1016/j.specom.2016.08.003
0167-6393/© 2016 Elsevier B.V. All rights reserved.

The approaches for QbE STD and QbE SDR that can be found in the literature differ in the amount of knowledge about the languages to be processed and in the need for modelling and resources. A search on speech system is language independent when no knowledge about the language of the documents is necessary, it being possible to perform the search without caring about the language. The use of cross-lingual approaches is common, especially when dealing with under-resourced languages, as they enable the use of models in one language when performing a search in spoken documents in a different language, with the obtaining of very promising results (Abad et al., 2013b; Rodriguez-Fuentes et al., 2014). According to the amount of modelling and resources, we can distinguish among ASR techniques, in which a speech recognizer in the target language is used to generate a textual or phonetic transcription of the documents; low-resource systems, in which some modelling is performed but does not require much knowledge about the data to be processed; and zero-resource systems, in which no modelling, resources or information about the language are necessary for performing the search. This latter approach has some advantages over the others since it does not require specific resources, language knowledge or modelling, but it has been less explored and the achieved results were not as good as those obtained with other approaches (Metze et al., 2013; 2012b).

Zero-resource QbE STD/SDR systems are usually based on pattern matching techniques: the spoken documents and the queries are represented by using a set of features, and a matching between the queries and the documents is performed in order to find occurrences of the queries in the documents, which is usually carried out by using the dynamic time warping (DTW) algorithm (Sakoe and Chiba, 1978) or any of its variants (Anguera, 2013; Anguera and Ferrarons, 2013; Mantena et al., 2014; Müller, 2007). The interest in zero-resource approaches for speech processing has recently increased, motivating the organisation of events such as the Zero Resource Speech challenge (Versteegh et al., 2015), which aim to discover linguistic units in an unsupervised way, but these approaches are still underdeveloped. In zero-resource search on speech, the use of Mel-frequency cepstral coefficients (MFCC) is as extended as in many other speech technologies tasks, either with or without some processing of the features (Calvo et al., 2014; Joder et al., 2012; Martinez et al., 2014). Some researchers experimented with other acoustic features such as the zero crossing rate, MPEG-7 low level descriptors (Vavrek et al., 2012), perceptual linear prediction (Carlin et al., 2011) or short-time frequency domain linear prediction features (Jansen et al., 2012). However, no features apart from the MFCCs were found to be notably useful for this task. Hence, which features to use in this task is still an open research problem. A widely used approach tackling this concern in other speech technologies tasks is the use of selection techniques on a large set of features. Unfortunately, the most common feature selection techniques used in pattern recognition tasks cannot be straightforwardly applied in this situation, since search on speech does not involve a division of the data to be processed into classes in order to be able to compute the relevance of the different features. In Lopez-Otero et al. (2015), an approach for feature selection in a low-resource QbE STD scenario was presented, and the experimental validation suggested that more is not always better, as results obtained when using a reduced set of features were better than when using the whole set.

In this study, we extend the work presented in Lopez-Otero et al. (2015) for feature selection in QbE STD/SDR tasks. The aforementioned method, which was applied to phoneme posteriorgram representations obtained by applying a phoneme recogniser, consisted in estimating the relevance of each phonetic unit in order to discard those that were not relevant for the scenario of application. To do so, the best alignment paths obtained from aligning the matching documents and queries in the training data were used to decompose the cost of each phonetic unit in order to obtain their relevance: to compute this relevance, it is assumed that the lower the cost of a phonetic unit in the best alignment path, the higher its relevance. In this paper, this focus is extended to feature selection in zero-resource scenarios and a novel method to estimate the relevance of the features is proposed. In this method, instead of only using the costs obtained in the best alignment path, we take into account the neighbouring region of this alignment path, assuming that it is preferable to have a “deep” best alignment path, i.e. to have a big variation of cost between the neighbouring region and the best alignment path, as this would indicate that a given feature serves to be as sure as possible about which path is the best one to choose.

Four different experimental frameworks were used to assess the performance of the proposed approach, namely those used for two QbE STD evaluations (Albayzin 2014 search on speech evaluation (Tejedor et al., 2014) and MediaEval spoken web search 2013 evaluation (Anguera et al., 2013)) and two QbE SDR evaluations (query-by-example search on speech task at MediaEval 2014 (Anguera et al., 2014a) and 2015 (Szöke et al., 2015)). The two feature selection techniques were compared with a dimensionality reduction approach based on principal component analysis (PCA), with the classic MFCC representation and with phoneme posteriorgram-based approaches. Using feature selection instead of PCA led to an overall superior performance, and the same happened when comparing with the low-resource strategies, which suggests that the performance gap between zero-resource techniques and more complex approaches is steadily decreasing. The experimental validation also showed that the use of a reduced set of features obtained a better performance than a brute force approach in which all the available features were used and, compared with the typical MFCC representation, a dramatic improvement was achieved.

The rest of this paper is organized as follows: the system that is used in this paper for QbE STD/SDR is described in Section 2; Section 3 describes the different strategies for feature selection in a QbE search on speech scenario; a summary of the experimental frameworks and metrics used to assess system performance are defined in Section 4; Section 5 describes the contrastive systems that are used in order to compare the proposed zero-resource approach with other strategies; Section 6 presents the experimental results; a discussion on the results is drawn in Section 7; and Section 8 summarizes some conclusions and future work.

2. Zero-resource search on speech approach

The zero-resource search on speech approach used in this paper consists of three stages: feature extraction, search and score normalization. The feature extraction step consists in extracting features from the waveforms of the queries and documents in order to have each spoken document represented by means of feature vectors, as described above. The rest of this Section describes the modification of DTW that was used in this work, namely sequential DTW (S-DTW), as well as the normalization techniques that were applied to the scores obtained from the search stage.

2.1. Feature extraction

There are two main approaches for feature extraction in the speech technologies literature: using a reduced set of features (usually MFCCs) against using large sets of features. The former approach aims at improving the computational efficiency of the system by representing speech using only those features that are relevant for the task, while the latter takes advantage of the powerful machines available nowadays in order to represent speech using as much information as possible. There is not an agreement in the research community about which of the two approaches is better, so