Speech Communication 103 (2018) 27–36

Contents lists available at ScienceDirect

Speech Communication

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

Phonetic subspace features for improved query by example spoken term detection

Dhananjay Ram[a,b,*], Afsaneh Asaei[a], Hervé Bourlard[a,b]

[a] Idiap Research Institute, Martigny, Switzerland
[b] Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 6 December 2017
Revised 25 April 2018
Accepted 13 July 2018
Available online 8 August 2018

Keywords:
Deep neural network
Phone posterior
Phonological posterior
Sparse representation
Dictionary learning
Query by example
Spoken term detection

ABSTRACT

This paper addresses the problem of detecting speech utterances from a large audio archive using a simple spoken query, hence referring to this problem as "Query by Example Spoken Term Detection" (QbE-STD). This still open pattern matching problem has been addressed in different contexts, often based on variants of the Dynamic Time Warping (DTW) algorithm. In the work reported here, we exploit Deep Neural Networks (DNN) and the so inferred phone posteriors to better model the phonetic subspaces and, consequently, improve the QbE-STD performance. Those phone posteriors have indeed been shown to properly model the union of the underlying low-dimensional phonetic subspaces. Exploiting this property, we investigate here two methods relying on sparse modeling and linguistic knowledge of sub-phonetic components. Sparse modeling characterizes the phonetic subspaces through a dictionary for sparse coding. Projection of the phone posteriors through reconstruction on the corresponding subspaces using their sparse representation enhance those phone posteriors. On the other hand, linguistic knowledge driven sub-phonetic structures are identified using phonological posteriors which consists of the probabilities of phone attributes estimated by DNNs, resulting in a new set of feature vectors. These phonological posteriors provide complementary information and a distance fusion method is proposed to integrate information from phone and phonological posterior features. Both posterior features are used for query detection using DTW and evaluated on AMI database. We demonstrate that the subspace enhanced phone posteriors obtained using sparse reconstruction outperforms the conventional DNN posteriors. The distance fusion technique gives further improvement in QbE-STD performance.

1. Introduction

Query by Example Spoken Term Detection (QbE-STD) refers to the task of detecting all audio documents from a database such that the documents contain a spoken query provided by a user. This enables the users to search over spoken audio archives using their own speech. The primary difference between QbE-STD and keyword spotting is that the user provides one or more examples of a spoken query instead of a textual query. In general, the query examples as well as test utterances can be spoken by different speakers in varying acoustic conditions without any constraints on the language and corresponding vocabulary. Since no training data is required nor provided, QbE-STD is a particular case of a zero-resource task.

A QbE-STD system is useful for searching through audio data generated by news channels, radio broadcasts, internet etc. These audio contents are produced everyday in multiple languages by a large number of diverse users. Due to the lack of knowledge about the language of interest and corresponding training data, it is difficult to build an automatic speech recognition (ASR) system and integrate it to a text based retrieval system to perform QbE-STD (Lee et al., 2015). Therefore, recent advances in QbE-STD are largely dominated by template matching techniques for its superior performance in zero-resource condition (Anguera et al., 2014; Rodriguez-Fuentes et al., 2014). The template based QbE-STD system primarily involves two steps: (1) extraction of feature vectors from the spoken query and the test audio, and (2) alignment of the query and test features using dynamic time warping (DTW) (Rabiner et al., 1978) or one of its variants (Müller, 2007; Zhang and Glass, 2009). Phone posterior features (posterior probabilities of a set of phonetic classes) (Hazen et al., 2009; Rodriguez-Fuentes et al., 2014) and bottleneck features (representation obtained from the bottleneck layer of a deep neural net-

* Corresponding author
E-mail addresses: dhananjay.ram@idiap.ch (D. Ram), afsaneh.asaei@idiap.ch (A. Asaei), herve.bourlard@idiap.ch (H. Bourlard).

https://doi.org/10.1016/j.specom.2018.07.001
0167-6393/

work) (Szöke et al., 2014; Chen et al., 2017) have been successful for QbE-STD. These features are extracted from deep neural networks (DNN) trained using multiple well-resourced languages. The bottleneck features can also be extracted in an unsupervised manner using labels generated from clustering techniques (Chen et al., 2016).

Our earlier attempts at QbE-STD rely on low-dimensional subspace structure of speech signal (Ram et al., 2015; 2016; 2017; 2018a). This structure of speech can be attributed to the constrained configuration of the human speech production system, leading to the generation of speech signals lying on low-dimensional, non-linear manifolds (Deng, 2004; King et al., 2007). The low-dimensional structure is exploited using sparse representation of speech data and QbE-STD is cast as a subspace detection problem between query and non-query speech (Ram et al., 2015; 2016; 2018a). This property of speech has also been exploited to perform robust speech recognition (Sainath et al., 2011; Gemmeke et al., 2011) as well as enhanced acoustic modeling (Dighe et al., 2016b). Our method presented a faster approach than template matching, however it lacked a framework to capture the temporal information inherent to speech. In contrast, we propose here to exploit the low-dimensional properties to obtain a better representation of the speech signal, before performing QbE-STD using DTW based template matching. In this way, we exploit the temporal information as well as low-dimensional structure of speech signal. To achieve this goal, we propose a data-driven and a knowledge-based approach to obtain better representation of speech and a fusion technique to combine information from different kinds of representations as discussed below.

(i) *Phonetic subspace representation - A data-driven approach* (Section 4): We propose to use sparse modeling as an unsupervised data-driven method to characterize the low-dimensional structures of sub-phonetic components (Elhamifar and Vidal, 2013; Rish and Grabarnik, 2014). To that end, we model the underlying phonetic subspaces using dictionary learning for sparse coding. The dictionaries are used to obtain sparse representation of the phone posteriors and we project them onto the phonetic subspaces through reconstruction. This approach leads to subspace enhanced phone posteriors such that the query and test posteriors are represented on a common subspace and reduces the effect of unstructured phonetic variations.

(ii) *Phonetic subspace representation - A knowledge-based approach* (Section 5): Alternative to the data-driven sparse modeling approach, we utilize linguistic knowledge for identifying the sub-phonetic attributes or phonological features (Chomsky and Halle, 1968). The phonological features are recognized as the atomic components of phone construction. The linguists define a binary mapping between the phone and phonological categories. We exploit DNN in probabilistic characterization of the phonological features, referred to as the *phonological posteriors* (Cernak et al., 2017). Due to the sub-phonetic nature of these features, they are less language dependent (Lee and Siniscalchi, 2013; Sahraeian et al., 2015) and can be helpful for a zero resource task like QbE-STD.

(iii) *Distance fusion* (Section 6): The proposed representations are exploited for QbE-STD using the DTW method presented in Rodriguez-Fuentes et al. (2014) (see Section 3 for details). To integrate the information from multiple feature representations, we propose to update the distance matrix for DTW by fusing the distances between the query and test utterance obtained from different kinds of feature representations. In contrast to Wang et al. (2013), we use non-uniform weights which are optimized using development queries.

The proposed methods are evaluated on two subsets of AMI database (IHM and SDM) with challenging conditions as presented in Section 8. The improvements obtained by our approach over the baseline system indicate the significance of subspace structure of speech for QbE-STD.

2. Related works

In this section, we summarize different techniques proposed for QbE-STD. The first set of methods consists of a two step approach: feature extraction and template matching as discussed earlier. The spoken queries as well as test utterances can be represented using mel frequency cepstral coefficient (MFCC) or perceptual linear prediction (PLP) based spectral features. These spectral features were initially investigated for template matching task (Sakoe and Chiba, 1978). However these features were outperformed by posterior features, which can be estimated from models trained in both supervised and unsupervised manner (Hazen et al., 2009; Rodriguez-Fuentes et al., 2014; Zhang and Glass, 2009). Gaussian mixture model (GMM) based posteriors are estimated from a GMM trained in an unsupervised manner where the feature dimensions correspond to posterior probabilities of different Gaussian components in the model (Zhang and Glass, 2009; Park and Glass, 2008). On the other hand, a deep boltzman machine (DBM) trained in unsupervised as well as semi-supervised manner can be used to extract posterior features. The unsupervised training of DBM can capture hierarchical structural information from unlabeled data. In Zhang et al. (2012), the authors first train a DBM using unlabeled data and then fine tune it using small amount of labeled data. In another approach, GMM based posteriors were used as labels for the DBM training (Zhang et al., 2012). Posteriors from DBM in both cases perform better than GMM posteriors for QbE-STD.

The supervised approach to extract posterior features primarily relies on training a DNN using labeled data. In case of zero resource languages, the DNN is first trained using data from different well resourced languages where the labels can indicate monophones, context dependent phones or senones (Hazen et al., 2009; Rodriguez-Fuentes et al., 2014). The DNN is then used to extract posterior features to perform template matching for QbE-STD. In this approach, the posteriors are interpreted as a characterization of instantaneous content of the speech signal, irrespective of the underlying language (Rodriguez-Fuentes et al., 2014). DNNs with bottleneck layer have also been trained in a similar multilingual setting to compute bottleneck features for QbE-STD (Szöke et al., 2014; Chen et al., 2017).

Features extracted from the spoken query and test utterance are used to compute a frame-level distance matrix and a DTW algorithm is used to find the degree of similarity between them. Standard DTW algorithm performs an end-to-end comparison between two temporal sequences, making it difficult to use for QbE-STD because the query can occur anywhere in the test utterance as a sub-sequence. In segmental DTW (Park and Glass, 2008), the distance matrix is segmented into overlapping diagonal bands where the width of the band indicates temporal distortion allowed for matching. But the width of each band limits its capability to deal with signals of widely varying speaking rate. Slope-constrained DTW (Zhang and Glass, 2009) was proposed to deal with this problem by penalizing the slope of warping path which maps the spoken query within a test utterance. It limits the number of frames to be mapped in the test audio corresponding to a frame in the query and vice versa. In sub-sequence DTW (Müller, 2007), the cost of insertion is forced to be 0 in the beginning and end of a query, which enables the warping path to begin and end at any point in the test audio and finds a sub-sequence best matching the query.

More recent approaches are aimed at minimizing the computational cost or memory footprints of the DTW based search