IDIAP RESEARCH REPORT

SPARSE SUBSPACE MODELING FOR QUERY BY EXAMPLE SPOKEN TERM DETECTION

Dhananjay Ram   Afsaneh Asaei   Hervé Bourlard

Idiap-RR-01-2016

JANUARY 2016

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

SPARSE SUBSPACE MODELING FOR QUERY BY EXAMPLE SPOKEN TERM DETECTION

Dhananjay Ram*†, *Afsaneh Asaei**, *Hervé Bourlard**†

*Idiap Research Institute, Martigny, Switzerland
†Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{dhananjay.ram,afsaneh.asaei,herve.bourlard}@idiap.ch

ABSTRACT

We cast the problem of query by example spoken term detection (QbE-STD) as subspace detection where query and background are modeled as a union of low-dimensional subspaces. The speech exemplars used for subspace modeling consist of class-conditional posterior probabilities obtained from deep neural network (DNN). The query and background training exemplars are exploited to model the underlying low-dimensional subspaces through dictionary learning and sparse coding. Given the dictionaries characterizing the query and background speech, QbE-STD amounts to subspace detection via sparse representation and the reconstruction error is used for binary classification. Furthermore, we rigorously investigate the relationship between the proposed method and the generalized likelihood ratio test. The experimental evaluation demonstrate that the proposed method is able to detect the query given a single exemplar and performs significantly better than one of the best QbE-STD baseline systems based on template matching.

Index Terms— Deep neural network posterior probabilities, Subspace sparse representation, Dictionary learning, Sparse modeling, Query by example spoken term detection.

1. INTRODUCTION

Query-by-example spoken term detection (QbE-STD) refers to the task of finding a spoken query within spoken audio. It enables voice search in the context of multi-lingual unconstrained audio data which can also be used for content indexing and retrieval applications.

1.1. Prior Works

A traditional QbE-STD approach is to convert spoken audio into a sequence of symbols and then perform text based search. In [1, 2, 3], the audio is first converted into a sequence of symbols using automatic speech recognition (ASR) and then lattice based search techniques are applied to detect the symbolic representation of the query. These techniques typically require large amount of transcribed data to train statistical acoustic model and language model for the underlying speech recognition system.

To apply the QbE-STD system on the raw data available on the web, it is important to process the data without any requirement for transcription. Hence, recent advances in QbE-STD are largely dominated by the exemplar-based template matching techniques for its superior performance to the statistical methods in low-resource condition [4, 5]. This approach is conducted in two steps. First, the query and test utterances are represented in terms of features or exemplars. The query and the test exemplars are then aligned using dynamic time warping (DTW) [6] or one of its variations [7]. The similarity of the query and test exemplars obtained from DTW are compared with a pre-defined threshold to find out possible regions of query occurrences. Both spectral and class-conditional posterior probabilities [8] are used as features to represent the audio exemplars. Although this approach requires a few query exemplars, it is sensitive to speaker and acoustic mismatch conditions. To overcome these limitations, model based approaches are being investigated [9]. In [10], acoustic units are discovered and modeled using HMM in an unsupervised manner. These units are then used to model the query and search for it in a test utterance.

1.2. Motivations and Contributions

This work is motivated by the success of exemplar-based sparse representation in classification and detection tasks [11, 12]. Previous studies in this direction are largely confined to the application of speech recognition and separation [13, 14, 15]. To the best of our knowledge, this has not been studied by other researchers in the context of spoken term detection. In contrast to the earlier work on exemplar-based sparse representation where spectral features are used as exemplars, we use DNN-based posterior features.

Speech utterances are a union of words which in turn consist of phonetic components and sub-phonetic attributes. Each linguistic component is produced using a few highly constrained articulatory mechanisms leading to generation of speech data living in a union of low-dimensional subspaces [16, 17, 18]. However, most existing speech classification and acoustic modeling methods do not explicitly consider this multi-subspace structure of speech. Sparse modeling is a promising technique to exploit this structure [19, 20].

Previously, in [21], we cast the query detection problem as the problem of subspace detection via sparse representation. To that end, a dictionary for characterizing the space of query and background exemplars was learned from training data at the *word* level. Namely, we assumed that the speech utterances are a set of *known* words. The individual word spaces were modeled as a union of low-dimensional subspaces learned through dictionary learning for sparse representation.

In this paper, we extend our preliminary work in several directions: Instead of word based dictionary learning, we use *phone* based dictionaries, hence, generalizing the applicability of our method for utterances composed of *unknown* words. We exploit the exclusiveness of query and background frames in a class-specific sparse recovery approach to improve the accuracy of query and background sparse representation. Moreover, we demonstrate how exploiting temporal information enables us to resolve the subspace intersection ambiguities due to the overlapping query and background phonetic components. Finally, we demonstrate the mathematical equivalence of the proposed method to the generalized likelihood ratio test for composite hypothesis testing. The experimental evaluation shows significant improvement over one of the strongest template matching techniques [4, 5].

2. SPARSE SUBSPACE POSTERIOR MODELING

In this section, we elaborate on modeling the query and background speech as a union of low-dimensional subspaces. The DNN class-conditional posterior probabilities are used as speech features for subspace modeling. The QbE-STD problem amounts to subspace detection where sparse recovery is applied to identify the underlying low-dimensional subspaces.

2.1. Union of Low-dimensional Subspaces

When speech is represented in terms of posterior features, the subspace corresponding to each linguistic class is a low-dimensional space [22]. Any speech utterance is comprised of individual linguistic classes, hence, it can be modeled as a union of low-dimensional subspaces. To state it more precisely, let $\mathcal{Q}$ and $\mathcal{B}$ denote the query and background manifold respectively such that

$$\mathcal{Q} = \cup_{i=1}^{m} \mathcal{Q}_i, \qquad \mathcal{B} = \cup_{i=1}^{n} \mathcal{B}_i \tag{1}$$

where $\{\mathcal{Q}\}_{i=1}^{m}$ and $\{\mathcal{B}\}_{i=1}^{n}$ are the corresponding constituent subspaces.

Any data point in a union of subspaces can be efficiently reconstructed by a combination of other points in the dataset [20]. Hence, to characterize the space of posterior features, all training exemplars should be considered. To alleviate the need of all training data, dictionary learning for sparse representation provides an unsupervised, yet effective, way of extracting an over-complete basis set to model the underlying subspaces. This approach reduces the computational cost and improves the accuracy of sparse modeling [23].

Given the dictionary which can characterize the underlying subspace model, the independent subspaces are guaranteed to be identified correctly using sparse recovery [20]. In the following Section 2.2, we explain how the query and background subspaces can be modeled using dictionary learning for sparse representation.

2.2. Query and Background Dictionaries

Dictionary learning refers to the task of learning an over-complete set of basis vectors from the training data such that the underlying subspaces can be represented as a sparse linear combination of these vectors. Given a training set of features $\mathbf{Y} = [\mathbf{y}_1,\ \mathbf{y}_2,\ ...,\ \mathbf{y}_T] \in \mathbb{R}^{K\times T}$, a dictionary $\mathbf{D} \in \mathbb{R}^{K\times M}$ and sparse representation $\mathbf{A} = [\boldsymbol{\alpha}_1,\ \boldsymbol{\alpha}_2,\ ...,\ \boldsymbol{\alpha}_T]$ for $\mathbf{Y}$; the objective function for dictionary learning algorithm is defined as

$$\arg\min_{\mathbf{D},\mathbf{A}} \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2}\|\mathbf{y}_t - \mathbf{D}\,\boldsymbol{\alpha}_t\|_2^2 + \lambda\|\boldsymbol{\alpha}_t\|_1\right) \tag{2}$$

where λ is the regularization parameter, K is the size of posterior feature vector, T is the number of training vectors for a given example, M is the number of basis vectors in the dictionary and $K \ll M$. The first term in this expression, quantifies the energy-based *reconstruction error*. The second term denotes the ℓ_1-norm of $\boldsymbol{\alpha}$ defined as $\|\boldsymbol{\alpha}\|_1 = \sum_i |\alpha_i|$ which controls the sparsity of $\boldsymbol{\alpha}_t$. The joint optimization of this objective function with respect to both $\mathbf{D}$ and $\boldsymbol{\alpha}_t$ simultaneously is non-convex, it can be solved as a convex objective by optimizing for one while keeping the other fixed [24].

In this work, we consider the fast online algorithm proposed by Mairal et al. [24] for its high performance in posterior based subspace modeling [23]. This algorithm is based on stochastic gradient descent optimization. It basically alternates between a step of sparse coding for the current training feature z_t and then optimizes the previous estimate of dictionary $\mathbf{D}^{(t-1)}$ to determine the new estimate $\mathbf{D}^{(t)}$ using stochastic gradient descent. The algorithm has been shortly summarized in Algorithm 1.

Algorithm 1 Online Dictionary Learning

```
Require: : $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T] \in \mathbb{R}^{K\times T}, \lambda \in \mathbb{R}$ : regularization
    parameter, initial estimate for dictionary $\mathbf{D}^{(0)} \in \mathbb{R}^{K\times M}$
for $t = 1$ to $T$ do
    Sparse Coding of $\mathbf{y}_t$ to determine $\alpha_t$:
        $\alpha_t = \arg\min_\alpha \left\{\frac{1}{2}\|\mathbf{y}_t - \mathbf{D}^{(t-1)}\alpha\|_2^2 + \lambda\|\alpha\|_1\right\}$
    Updating $\mathbf{D}^{(t)}$ with $\mathbf{D}^{(t-1)}$ as warm restart:
        $\mathbf{D}^{(t)} = \arg\min_\mathbf{D} \left\{\frac{1}{t}\sum_{i=1}^{t}(\frac{1}{2}\|\mathbf{y}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda\|\alpha_i\|_1)\right\}$
end for
return $\mathbf{D}^{(T)}$
```

In this paper, we learn two sets of dictionaries separately:

1. The query dictionary is learned from query exemplars and denoted by $\mathbf{D}_\mathcal{Q}$.
2. The background dictionary includes the set of phone-specific dictionaries learned from training exemplars of phone posterior features and denoted by $\mathbf{D}_\mathcal{B} = \{\mathbf{D}_1 \ldots \mathbf{D}_p \ldots \mathbf{D}_P\}$ where p indicates different phones.

These two classes will always have shared phonetic components which makes the classification a difficult task. The primary discriminating property between the two classes is the temporal information inherent to query which is modeled in the query dictionary. On the other hand, there is no such structure present in the background dictionary due to separate phone-based dictionaries. To exploit the temporal structure, a sequence of c frames is concatenated as $\tilde{\mathbf{y}} = [\mathbf{y}_{t-c}^\top \cdots \mathbf{y}_t^\top \cdots \mathbf{y}_{t+c}^\top]^\top$ and form the input feature used for dictionary learning and sparse representation. This mechanism is referred to as *context appending* which is also a typical approach to incorporate the dynamics of the features [23].

2.3. Spoken Term Subspace Detection

The underlying low-dimensional subspaces of speech posterior features can be identified using sparse representation. Given a test posterior vector $\mathbf{z}_t$ and the query and background dictionaries $\mathbf{D}_\mathcal{Q}$ and $\mathbf{D}_\mathcal{B}$, the test vector can be represented as a sparse linear combination of dictionary atoms characterizing the space of query or background manifolds. Given the over-complete dictionaries, the query and background sparse representations of a posterior feature vector $\mathbf{z}_t$ is obtained by the following optimization problems:

$$\boldsymbol{\alpha}_t^\mathcal{Q} = \arg\min_{\boldsymbol{\alpha}} \left\{\frac{1}{2}\|\mathbf{z}_t - \mathbf{D}_\mathcal{Q}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1\right\} \tag{3}$$

$$\boldsymbol{\alpha}_t^\mathcal{B} = \arg\min_{\boldsymbol{\alpha}} \left\{\frac{1}{2}\|\mathbf{z}_t - \mathbf{D}_\mathcal{B}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1\right\} \tag{4}$$

The coefficients of the sparse representation of $\mathbf{z}_t$ in the query and background spaces are $\boldsymbol{\alpha}_t^\mathcal{Q}$ and $\boldsymbol{\alpha}_t^\mathcal{B}$ respectively. The reconstructed vector using the corresponding sparse coefficient vectors will be,

$$\hat{\mathbf{z}}_t^\mathcal{Q} = \mathbf{D}_\mathcal{Q}\boldsymbol{\alpha}_t^\mathcal{Q}, \qquad \hat{\mathbf{z}}_t^\mathcal{B} = \mathbf{D}_\mathcal{B}\boldsymbol{\alpha}_t^\mathcal{B}$$

The subspace which can best represent a test vector $\mathbf{z}_t$ corresponds to the least reconstruction error. Hence, we use the

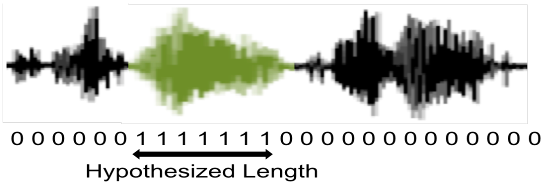


Fig. 1. Hypothesized length of the query in a test utterance.

Euclidean-norm based reconstruction error to perform binary classification. The reconstruction errors are calculated as follows

$$e^{\mathcal{Q}}(\mathbf{z}_t) = \|\mathbf{z}_t - \hat{\mathbf{z}}_t^{\mathcal{Q}}\|_2 = \|\mathbf{z}_t - \mathbf{D}^{\mathcal{Q}}\boldsymbol{\alpha}_t^{\mathcal{Q}}\|_2 \tag{5}$$

$$e^{\mathcal{B}}(\mathbf{z}_t) = \|\mathbf{z}_t - \hat{\mathbf{z}}_t^{\mathcal{B}}\|_2 = \|\mathbf{z}_t - \mathbf{D}^{\mathcal{B}}\boldsymbol{\alpha}_t^{\mathcal{B}}\|_2 \tag{6}$$

where, $e^{\mathcal{Q}}(\mathbf{z}_t)$ and $e^{\mathcal{B}}(\mathbf{z}_t)$ are the error terms corresponding to the query and background classes. The errors are then used to take a frame-level decision by calculating their difference as

$$\Delta(\mathbf{z}_t) = e^{\mathcal{B}}(\mathbf{z}_t) - e^{\mathcal{Q}}(\mathbf{z}_t) \tag{7}$$

which is compared with a pre-defined threshold δ. If $\Delta(\mathbf{z}_t) > \delta$, $\mathbf{z}_t$ is labeled as a query-frame, otherwise $\mathbf{z}_t$ is marked as a background-frame. The frame-level decisions are then accumulated to form an utterance level decision and to detect whether the query occurs once or more in the test utterance. This is done by counting the continuous number of frames detected as the query. This provides us with the hypothesized length of the query in a test utterance. Fig. 1 depicts this procedure to obtain the hypothesized length. This length is compared with a pre-calculated threshold to take the final decision [25]. Although the frame-level processing is not able to exploit the temporal information inherent in speech, this information is captured through context appending as discussed in Section 2.2 to obtain the frame-level decisions. Although very simple (and disregarding temporal information), this decoding procedure has been shown advantageous to the Viterbi algorithm in the framework of hidden Markov model (HMM) for keyword detection task [25]. We will see in Section 3.4 that this decision making approach is effective and outperforms an efficient DTW-based baseline system in QbE-STD evaluations.

2.4. Relation to Generalized Likelihood Ratio Test

The proposed approach is closely related to the generalized likelihood ratio test for composite hypothesis testing. We assume that each test exemplar is modeled as $\mathbf{z}_t = \mathbf{D}\boldsymbol{\alpha}_t + \mathbf{n}_t$ where $\mathbf{D}$ is an over-complete dictionary and $\boldsymbol{\alpha}_t$ is a sparse latent variable with Laplace prior distribution

$$p(\boldsymbol{\alpha}_t) \sim \left(\frac{\lambda}{2}\right)^M \exp(-\lambda\|\boldsymbol{\alpha}\|_1). \tag{8}$$

with a parameter $\lambda > 0$. We assume the model mismatch $\mathbf{n}_t$ to be an independent Gaussian noise distributed as $\mathcal{N}(0, \sigma^2\mathbf{I})$. Hence, the distribution of a test exemplar $\mathbf{z}_t$ given the latent variable $\boldsymbol{\alpha}_t$ is given by:

$$p(\mathbf{z}_t|\boldsymbol{\alpha}_t = \boldsymbol{\alpha}; \mathbf{D}) \sim \mathcal{N}(\mathbf{D}\boldsymbol{\alpha}, \sigma^2\mathbf{I}) \tag{9}$$

For each test posterior, we define the composite hypothesis testing problem as

$$\begin{aligned} H_0 &: \ \mathbf{z}_t = \mathbf{D}_{\mathcal{Q}}\,\boldsymbol{\alpha}_t + \mathbf{n}_t \\ H_1 &: \ \mathbf{z}_t = \mathbf{D}_{\mathcal{B}}\,\boldsymbol{\alpha}_t + \mathbf{n}_t, \end{aligned} \tag{10}$$

The maximum likelihood estimate of $\boldsymbol{\alpha}_t$ is obtained as

$$\arg\max_{\boldsymbol{\alpha}} p(\boldsymbol{\alpha}|\mathbf{z}_t; \mathbf{D}) = \arg\max_{\boldsymbol{\alpha}} p(\mathbf{z}_t|\boldsymbol{\alpha}; \mathbf{D})p(\boldsymbol{\alpha}) \tag{11}$$

Substituting (8) and (9), the maximum likelihood estimate of $\boldsymbol{\alpha}_t$ amounts to (3) and (4) if $\mathbf{D}$ is chosen as either the query $\mathbf{D}_{\mathcal{Q}}$ or background $\mathbf{D}_{\mathcal{B}}$ dictionary. Hence, the generalized likelihood ratio test becomes

$$\frac{p(\mathbf{z}_t; \boldsymbol{\alpha}_t^{\mathcal{Q}}|H_0)}{p(\mathbf{z}_t; \boldsymbol{\alpha}_t^{\mathcal{B}}|H_1)} = \frac{\|\mathbf{z}_t - \mathbf{D}_Q\,\boldsymbol{\alpha}_t^{\mathcal{Q}}\|_2}{\|\mathbf{z}_t - \mathbf{D}_B\,\boldsymbol{\alpha}_t^{\mathcal{B}}\|_2} \underset{H_1}{\overset{H_0}{\lessgtr}} \delta' \tag{12}$$

which leads to a solution equivalent to (7).

3. EXPERIMENTAL ANALYSIS

This section describes the spoken term detection experiments conducted to analyse and evaluate the performance of the proposed sparse subspace modelling method.

3.1. Database and Speech Exemplars

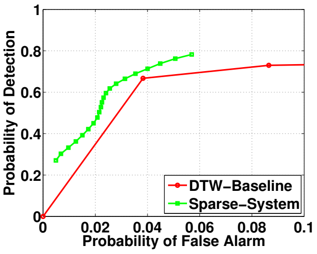
We use Numbers'95 database [26] for our experiments. In total, there are 31 numbers spoken in form of continuous speech, out of which 11 words are used for detection experiments. These numbers are 'zero' through 'nine' and 'oh'. There are around 17 k sentences in the database, among which 60% are used for training, 20% for development and the rest for testing. The training data is used to train a DNN which computes the phone posterior probabilities. More details about DNN training may be found in [23]. Indeed, the DNN can be learned using any well resourced speech database with a standard phone set. The DNN is used to extract the phone posterior features from the utterances. The query examples are chosen randomly from the training database to model the query. The same set of examples are used in different systems.

3.2. Baseline System

We have chosen the DTW based query by example spoken term detection system presented in [5] as our baseline system. As a brief overview of the system, it applies DTW matching of reference query with the test utterance in a recursive manner. In the first pass, the system hypothesizes a detected region with corresponding score. Then, the system searches again in the non-hypothesized region given the following three conditions are satisfied: (1) the score of the current hypothesis is greater than a given threshold T, (2) the non-hypothesized speech segment has long enough duration (half the query length) and (3) the number of detections (already hypothesized + currently computed) is less than a given threshold M. For our experiments, we have considered the same thresholds as [5], i.e. T = 0.85, and M = 7. If more than one example of the same query is given, we have used DTW matching to map the frames and averaged the matched frames to generate an average reference query [5, 27].

3.3. Sparse Subspace Detection

The first step is to learn the dictionaries for query and background classes. The query dictionaries are learned from the given examples of the query. The background dictionary includes all phone-specific dictionaries. The phone dictionaries can be learned from the same data used to train the DNN for feature extraction. We have used phone data from the training set of Numbers'95 database. Thus, we have 27 different phone dictionaries including a silence phone. We use the query and background (phone) dictionaries independently



(a) 1 Example

(b) 10 Examples

Fig. 2. ROC curves for the proposed sparse subspace detection system and the baseline DTW system [5] using different numbers of query examples available.

as (3) and (4) for sparse recovery of a test frame. There are 27 different reconstruction errors corresponding to the individual phone dictionaries, and the minimum error is used to calculate $\Delta(\mathbf{z}_t)$ in (7). Relying on the fact that a test frame can only be associated to a single phone subspace, choosing the minimum reconstruction error leads to selection of the most competitive underlying subspace of the background.

The query and background reconstruction errors or $\Delta(\mathbf{z}_t)$ in (7) are used to take frame-level binary decision. The frame-level decisions are then counted to estimate the length of a hypothesized query. Final decision is made by comparing the hypothesized length to half of the (average) query length.

3.4. QbE-STD Performance

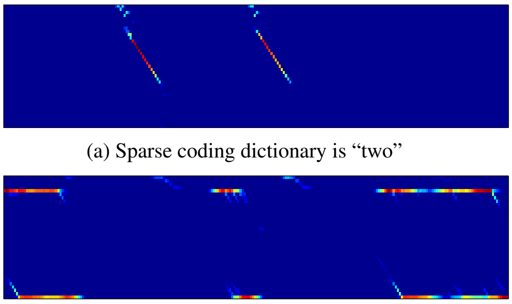
The receiver operating characteristic (ROC) curves are computed by varying the frame-level threshold δ in a predefined range for both sparse subspace detection as well as DTW based systems and illustrated in Fig. 2. Similar to the DTW system, half of the (average) query length is used as the minimum hypothesized length. The results are averaged for all digits used as the query.

We consider two cases where only a single query example (Fig. 2-(a)) and 10 examples (Fig. 2-(b)) are provided. In the case of single query example, the query dictionary $\mathbf{D}_Q$ consists of the query exemplars. In the case of 10 query examples, one of them is used for initializing the dictionary whereas the rest are used for dictionary learning (2). The value of λ is optimized over the development data as 0.8. The number of frames used for context appending is 17 corresponding to the context size of $c = 8$.

We can see that increasing the number of query examples improves the query subspace modeling thus enhancing the performance of the proposed method. Incorporating new examples of the query is more effective for sparse subspace detection compared to the baseline DTW system. It can be explained by different methodology adopted to incorporate the new data information. Clearly, dictionary learning captures more information from the examples compared to template averaging [5, 27] and provides a better characterization of the query space.

3.5. Sequential Structure

The temporal dependency of adjacent frames can be captured while learning the query dictionary via context appending as discussed in Section 2.2. This structure can be embedded in dictionary design once a sequence of context appended query exemplars is used for initialization of the dictionary learning algorithm. Due to sequencing information at initialization, the resulting dictionary atoms will represent a sequential basis set for query characterization. Sparse recovery of the test utterance using the query dictionary activates the atoms in a sequential manner in the region where the query occurs, leading to a sequential structure underlying the sparse coefficients. This property is illustrated in Fig. 3.



(a) Sparse coding dictionary is “two”

(b) Sparse coding dictionary is “nine”

Fig. 3. Sparse representation matrices obtained after decoding the utterance “three-two-eight-two-four”. Assuming that the query is “two”, the sparse coefficients exhibit a sequential structure when the dictionary of word “two” is used for sparse recovery whereas this structure is missing when a wrong dictionary “nine” is used.

The test utterance in this example contains the word sequence: “three-two-eight-two-four”; the sparse representation matrix depicted in Fig. 3-(a) is obtained using the dictionary of “two”. The sequential structure can be seen at the region of occurrences of “two”. On the other hand, if we use dictionary of “nine”, this sequential structure is no longer available as depicted in Fig. 3-(b). This structure can be exploited to devise a structured sparse recovery algorithm [12].

4. CONCLUSIONS

We proposed a novel spoken term detection approach based on sparse representation of the posterior exemplars. In contrast to the state of the art template matching methods, we presented the problem as subspace detection where the query and background are modeled using dictionary learning and sparse coding techniques. Sparse representation of the test posterior exemplars using the dictionaries characterizing the space of query and background leads to discrimination of the underlying subspaces thus enables classification via reconstruction error. Query decision is then simply performed by accumulating frame basis decisions (frames belonging to the query) over the hypothesized template, resulting in a very simple decoding process. However, in spite of this simplicity (and potential for improvements), the proposed approach has been shown here to outperform one of the best DTW baseline systems, demonstrating the great potential of our sparse subspace detection method. We plan to learn the universal phone dictionaries to evaluate the system on multilingual zero resourced spoken term detection tasks. To that end, discriminative dictionary learning techniques can be considered to alleviate the ambiguities of the intersecting (shared) phonetic subspaces. Furthermore, dictionary learning can be performed at any sub-phonetic levels suitable for speech representation.

5. ACKNOWLEDGMENTS

The research leading to these results has received funding from by SNSF projects on “Parsimonious Hierarchical Automatic Speech Recognition (PHASER)” grant agreement number 200021-153507 and “Adaptive Multilingual Speech Processing (A-MUSE)” grant agreement number 200020-144281.

6. REFERENCES

[1] Wade Shen, Christopher M White, and Timothy J Hazen, "A comparison of query-by-example methods for spoken term detection," Tech. Rep., DTIC Document, 2009.

[2] Carolina Parada, Abhinav Sethy, and Bhuvana Ramabhadran, "Query-by-example spoken term detection for OOV terms," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 404–409.

[3] I Szoke, M Fapso, Lukáš Burget, and J Cernocky, "Hybrid wordsubword decoding for spoken term detection," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Citeseer, 2008, pp. 42–48.

[4] Xavier Anguera, Luis J Rodriguez-Fuentes, Igor Szoke, Andi Buzo, Florian Metze, and Mikel Penagarikano, "Query-by-example spoken term detection evaluation on low-resource languages," in *The 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'14)*, 2014.

[5] Luis Javier Rodriguez-Fuentes, Amparo Varona, Mike Penagarikano, Germán Bordel, and Mireia Diez, "High-performance query-by-example spoken term detection on the sws 2013 evaluation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7819–7823.

[6] Lawrence R Rabiner, Aaron E Rosenberg, and Stephen E Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," *The Journal of the Acoustical Society of America*, vol. 63, no. S1, pp. S79–S79, 1978.

[7] Yaodong Zhang and James R Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 398–403.

[8] Timothy J Hazen, Wade Shen, and Christopher White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 421–426.

[9] Chun-an Chan and Lin-shan Lee, "Model-based unsupervised spoken term detection with spoken queries," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 7, pp. 1330–1342, 2013.

[10] Chia-ying Lee and James Glass, "A nonparametric bayesian approach to acoustic model discovery," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 40–49.

[11] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma, "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210–227, 2009.

[12] Yi Chen, Nasser M Nasrabadi, and Trac D Tran, "Sparse representation for target detection in hyperspectral imagery," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 3, pp. 629–640, 2011.

[13] Tara N Sainath, Bhuvana Ramabhadran, Michael Picheny, David Nahamoo, and Dimitri Kanevsky, "Exemplar-based sparse representation features: From TIMIT to LVCSR," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 8, pp. 2598–2613, 2011.

[14] Jort F Gemmeke, Tuomas Virtanen, and Antti Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2067–2080, 2011.

[15] Afsaneh Asaei, *Model-based Sparse Component Analysis for Multiparty Distant Speech Recognition*, Ph.D. thesis, École Polytechnique Fédéral de Lausanne (EPFL), 2013.

[16] Li Deng, "Switching dynamic system models for speech articulation and acoustics," in *Mathematical Foundations of Speech and Language Processing*, pp. 115–133. Springer New York, 2004.

[17] Leo J Lee, Paul Fieguth, and Li Deng, "A functional articulatory dynamic model for speech production," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2001, vol. 2, pp. 797–800.

[18] Afsaneh Asaei, Milos Cernak, and Hervé Bourlard, "On compressibility of neural network phonological features for low bit rate speech coding," in *Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015.

[19] Yonina C Eldar and Moshe Mishali, "Robust recovery of signals from a structured union of subspaces," *Information Theory, IEEE Transactions on*, vol. 55, no. 11, pp. 5302–5316, 2009.

[20] Ehsan Elhamifar and Rene Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 11, pp. 2765–2781, 2013.

[21] Dhananjay Ram, Afsaneh Asaei, Pranay Dighe, and Hervé Bourlard, "Sparse modeling of posterior exemplars for keyword detection," in *Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015.

[22] Pranay Dighe, Gil Luyet, Afsaneh Asaei, and Hervé Bourlard, "Exploiting low-dimensional structures to enhance dnn based acoustic modeling in speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2016.

[23] Pranay Dighe, Afsaneh Asaei, and Hervé Bourlard, "Sparse modeling of neural network posterior probabilities for exemplar-based speech recognition," *Speech Communication*, 2015.

[24] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research (JMLR)*, vol. 11, pp. 19–60, 2010.

[25] Hamed Ketabdar, Jithendra Vepa, Samy Bengio, and Hervé Bourlard, "Posterior based keyword spotting with a priori thresholds," in *International Conference on Spoken Language Processing (ICSLP)*, 2006.

[26] Ronald A Cole, Mike Noel, Terri Lander, and Terry Durham, "New telephone speech corpora at cslu.," in *Eurospeech*. Citeseer, 1995.

[27] Guoguo Chen, Carolina Parada, and Tara N Sainath, "Query-by-example keyword spotting using long short-term memory networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.