IDIAP RESEARCH REPORT

SUBSPACE DETECTION OF DNN POSTERIOR PROBABILITIES VIA SPARSE REPRESENTATION FOR QUERY BY EXAMPLE SPOKEN TERM DETECTION

Dhananjay Ram Afsaneh Asaei Hervé Bourlard

Idiap-RR-06-2016

APRIL 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

Subspace Detection of DNN Posterior Probabilities via Sparse Representation for Query by Example Spoken Term Detection

Dhananjay Ram[1,2]*, Afsaneh Asaei*[1]*, Hervé Bourlard*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{dhananjay.ram,afsaneh.asaei,herve.bourlard}@idiap.ch

Abstract

We cast the query by example spoken term detection (QbE-STD) problem as subspace detection where query and background subspaces are modeled as union of low-dimensional subspaces. The speech exemplars used for subspace modeling are class-conditional posterior probabilities estimated using deep neural network (DNN). The query and background training exemplars are exploited to model the underlying low-dimensional subspaces through dictionary learning for sparse representation. Given the dictionaries characterizing the query and background subspaces, QbE-STD is performed based on the ratio of the two corresponding sparse representation reconstruction errors. The proposed subspace detection method can be formulated as the generalized likelihood ratio test for composite hypothesis testing. The experimental evaluation demonstrate that the proposed method is able to detect the query given a single example and performs significantly better than a highly competitive QbE-STD baseline system based on dynamic time warping (DTW) for exemplar matching.

Index Terms: Deep neural network posterior probabilities, Subspace detection, Dictionary learning, Sparse representation.

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 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 matching techniques for its superior performance to the statistical methods in low-resource conditions [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, 5]. 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 speech exemplars. Although this approach requires a few query examples, it is sensitive to speaker and acoustic mismatch conditions. To overcome these limitations, model based approaches have been investigated [9]. In [10], new acoustic units are discovered and modeled using hidden Markov model (HMM) in an unsupervised manner. These units are then used to model the query for model-based query detection.

1.2. Our Contributions

We propose to cast the query detection problem as subspace detection where query and background subspaces are modeled through dictionary learning for sparse representation. This idea is motivated by the success of exemplar-based sparse representation in classification and detection tasks [11, 12]. In the context of speech processing, it has been widely studied for robust speech recognition [13, 14, 15]. We aim to study this approach for QbE-STD application. In contrast to the earlier work on exemplar-based sparse representation where spectral features are used as exemplars, we use DNN posterior probabilities as speech exemplars.

In our preliminary development of posterior sparse representation for keyword detection, we assumed that background speech consists of known words [16]. Hence, the background subspace was modeled as a union of low-dimensional word subspaces using dictionary learning with word exemplars. In this paper, we extend our initial framework in several directions. Instead of word based dictionary learning, we use phone based dictionaries, thus, generalizing the applicability of our method for utterances composed of unknown words. Since the background and query may have shared phonetic components, a large temporal context is exploited through appending adjacent posterior vectors to form contextually rich exemplars for dictionary learning. We present experimental evaluation of the proposed method based on AMI meeting corpus [17]. The sparse subspace detection is shown to be equivalent to the generalized likelihood ratio test for composite hypothesis testing where the posterior exemplars admit sparse representation in an over-complete dictionary and the residual error is assumed to have Gaussian distribution.

In the rest of the paper, subspace modeling and detection of query and background posteriors are described in Sections 2 and 3 respectively. Experimental results are presented in Section 4 and the conclusions are drawn in Section 5.

2. Subspace Modeling

In this section, we elaborate on the procedure to characterize the space of query and background posterior exemplars as union of low-dimensional subspaces.

2.1. Union of Low-dimensional Subspaces

When speech is represented in terms of posterior probabilities, the subspace corresponding to each sub-word class is a low-dimensional space [18, 19]. Accordingly, a speech utterance comprised of sub-word classes, 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 space 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 $\{\mathcal{Q}_i\}_{i=1}^{m}$ and $\{\mathcal{B}_i\}_{i=1}^{n}$ are the corresponding constituent subspaces, and m and n denote the number of composing sub-word classes respectively.

Any data point in union of low-dimensional subspaces can be efficiently reconstructed by a sparse combination of other points in that space, a property referred to as the self-expressiveness [20]. To alleviate the need of all training data, dictionary learning for sparse representation provides an 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 [21].

Given the dictionary for characterizing the underlying subspaces, the independent subspaces are guaranteed to be identified correctly using sparse representation [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 vectors from the training exemplars such that each training exemplar can be reconstructed as a sparse linear combination of the dictionary vectors (atoms).

Denoting a training set of T training exemplars with $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T\}$, and their sparse representations using dictionary $\mathbf{D} \in \mathbb{R}^{K \times M}$ with $\mathbf{A} = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_T\}$, where K is the dimension of exemplar vectors, and M is the number of dictionary atoms, the objective function for dictionary learning 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. The first term in this expression, quantifies the *reconstruction error*. The second term denotes the ℓ_1-norm of $\boldsymbol{\alpha}$ defined as $\|\boldsymbol{\alpha}\|_1 = \sum_i |\alpha_i|$ which quantifies 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 [22].

In this work, we consider the fast online algorithm proposed by Mairal et al. [22] for its good performance in posterior based subspace modeling [21]. This algorithm is based on stochastic gradient descent optimization. It basically alternates between a step of sparse representation for the current training feature $\mathbf{y}_t$ 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 is summarized in Algorithm 1.

Algorithm 1 Online Dictionary Learning

Require: : $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}, \lambda, \mathbf{D}^{(0)}$ (initialization)
1: **for** $t = 1$ to T **do**
2: Sparse representation of $\mathbf{y}_t$ to determine $\boldsymbol{\alpha}_t$:
$\boldsymbol{\alpha}_t = \arg\min_{\boldsymbol{\alpha}} \left\{ \frac{1}{2} \|\mathbf{y}_t - \mathbf{D}^{(t-1)}\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1 \right\}$
3: Updating $\mathbf{D}^{(t)}$ with $\mathbf{D}^{(t-1)}$ as warm restart:
$\mathbf{D}^{(t)} = \arg\min_{\mathbf{D}} \left\{ \frac{1}{t} \sum_{i=1}^{t} (\frac{1}{2} \|\mathbf{y}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \|\boldsymbol{\alpha}_i\|_1) \right\}$
4: **end for**
5: **return** $\mathbf{D}^{(T)}$

We learn different dictionaries to model the query and background subspaces:

1. The query dictionary denoted by $\mathbf{D}_{\mathcal{Q}}$.
2. The background dictionary denoted by $\mathbf{D}_{\mathcal{B}}$.

The query dictionary is learned from the available query examples. The background dictionary must represent any speech other than the query. However, the absence of lexical information corresponding to the query prevents us form constructing a very discriminative background dictionary. Hence, we propose to design a universal background dictionary for all potential queries. Such a background dictionary includes the set of all phone-specific dictionaries as $\mathbf{D}_{\mathcal{B}} = \{\mathbf{D}_1, \ldots, \mathbf{D}_P\}$ where P indicates the number of phones. The phone-specific dictionaries can be learned from a well-resourced speech database.

The query and background dictionaries represent some shared phonetic components which makes classification a difficult task. The primary discriminating property between the two classes is the temporal information underlying the query exemplars which is modeled in the query dictionary. On the other hand, there is no such structure present in the background dictionary due to modeling the separate phone-based dictionaries.

To exploit the temporal structure, a sequence of c posterior exemplars are concatenated as $\tilde{\mathbf{y}}_t = [\mathbf{y}_{t-c}^\top \ldots \mathbf{y}_t^\top \ldots \mathbf{y}_{t+c}^\top]^\top$, and form a contextually rich exemplar for dictionary learning and sparse representation. This mechanism is referred to as *context appending* which is a typical approach to incorporate the dynamics of the exemplars [14, 21].

3. Subspace Detection

Once the query and background subspaces are modeled, the QbE-STD problem is cast as subspace detection where the reconstruction errors of the respective sparse representations are used to detect the underlying subspace.

3.1. Detection via Sparse Reconstruction Errors

Given a test posterior exemplar $\mathbf{z}_t$ and the query and background dictionaries $\mathbf{D}_{\mathcal{Q}}$ and $\mathbf{D}_{\mathcal{B}}$, the test exemplar can be represented as a sparse linear combination of dictionary atoms characterizing the query or background subspaces. Given the over-complete dictionaries, the query and background sparse representations of a posterior exemplar $\mathbf{z}_t$ is obtained through 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 $\mathbf{z}_t$ over the query and background dictionaries are denoted by $\boldsymbol{\alpha}_t^{\mathcal{Q}}$ and $\boldsymbol{\alpha}_t^{\mathcal{B}}$ respectively. The reconstructed vectors using the corresponding sparse representations 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 Euclidean-norm based reconstruction error to perform binary classification [11]. The reconstruction errors are calculated as follows

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

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

The errors are then used to take a frame-level decision by calculating their difference as

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

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 labeled as a background-frame. The frame-level decisions are then accumulated to form an utterance level decision and to detect whether the query occurs in the test utterance. This is achieved by counting the continuous number of frames detected as the query. It provides us with the hypothesized length of the query in a test utterance [23]. Figure 1 depicts this procedure to obtain the hypothesized length. This length is compared with a pre-calculated threshold to take the final decision.

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. In spite of being simple (and disregarding temporal information), this decoding procedure has been found advantageous to the Viterbi algorithm in the framework of hidden Markov model (HMM) for keyword detection task [23]. We will see in Section 4.4 that this decision making approach is indeed effective, and outperforms a highly competitive DTW-based baseline system in QbE-STD evaluation [4, 5].

3.2. 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

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

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:

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

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

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

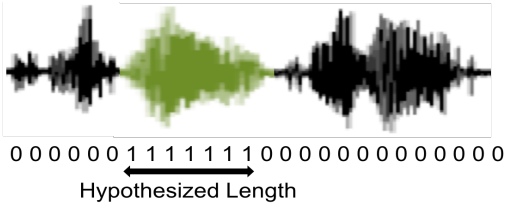

Figure 1: *Hypothesized length of the query in a test utterance.*

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

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

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

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

which leads to a solution equivalent to (7).

4. Experimental Analysis

The experiments are conducted to evaluate the performance of the proposed subspace detection method in challenging scenarios when very few or just one query example is provided for QbE-STD, and the query and background are conversational spontaneous speech with competitive speakers.

4.1. Data

The AMI meeting corpus [17] is used for the experiments where the training, development and evaluation sets are as [24]. That gives us about 81 hours for DNN training and about 9 hours for each of the development and evaluation data. Although, the meeting language was English, many participants were non-native speakers. Also, the headset recordings contain considerable amount of overlapping speech (competing speakers). There are approximately 12k words in the training, out of which 50 most frequent words are used for our detection experiments including very short words such as "my" to long words such as "television". The evaluation set is used for QbE-STD evaluation.

We have used Kaldi toolkit [25] to train a DNN for estimation of mono-phone posterior probabilities. The CMU pronunciation dictionary[1] is used for lexical modeling, which consist of 39 non-silence phones. Additionally, 4 phones are considered for silence. Hence, the phone posterior features are 43 dimensional. The query examples are chosen randomly from the training set to model the query. The same set of examples are used to evaluate the baseline and the proposed system.

4.2. Baseline System

The DTW based QbE-STD system presented in [5] is used as a highly competitive baseline system [4]. This method 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 s, (2) the non-hypothesized speech segment has long enough duration (half

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

the query length) and (3) the number of detections (already hypothesized + currently computed) is less than a given threshold N. For our experiments, $N = 7$ is optimized over the development set. 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, 26].

4.3. Subspace Modeling and Detection

The first step is to construct 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 to learn the corresponding phone dictionaries.

We use the query and background phone dictionaries independently as (3) and (4) for sparse representation of a test frame. There are different reconstruction errors corresponding to the different phone dictionaries, and the average error is used as the background score.

Alternative to the average reconstruction error of the background phone dictionaries, the minimum of them can also be used as the background score [16, 27]. However, we found that the average score yields better detection performance. We speculate that the reason is due to the difficulty in dictionary learning using a few random examples of the AMI corpus. Since this corpus has many occurrences of distorted speech due to non-native speaking and competing talkers which are represented in the query dictionary, the reconstruction error is not a score competitive to the well-trained background dictionaries. Hence, the average background score is a more conservative choice that leads to higher detection rate without increasing the probability of false alarm.

The difference of 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 pulled to estimate the length of a hypothesized query. Final decision is made by comparing the hypothesized length to the average query length (c.f. Figure 1).

4.4. QbE-STD Performance

The receiver operating characteristic (ROC) curves are computed by varying the thresholds δ and s in predefined ranges for sparse subspace detection and DTW based systems. For both systems, the average query length is used as the minimum hypothesized length. The results are averaged over all words used as the query to obtain the final ROC as illustrated in Figure 2.

We consider two cases where only a single query example and 10 examples are provided. In the case of single query example, the query dictionary $\mathbf{D}_{\mathcal{Q}}$ consists of the only query example. In the case of 10 query examples, one of them is used for initializing the dictionary whereas the rest are used for learning the dictionary. The value of different parameters are optimized over the development set. Clearly, in both cases our proposed system performs significantly better than the baseline system.

Unlike the DTW baseline system, the sparse subspace detection does not show much improvement when 10 examples are provided. This observation is opposite to the results previously obtained on a clean (simple) database [27] where incorporating more examples of the query were found more effective for the sparse method compared to the baseline DTW system. This issue can be attributed to the large variability and overlap present in the utterances of AMI corpus. When the query examples are noisy, the effect of noise is averaged out through the query averaging for DTW whereas the noisy examples are represented in the dictionary learning. On the contrary, when the query examples are clean, dictionary learning is more effective to characterize the variability present in the data than the simple averaging method.

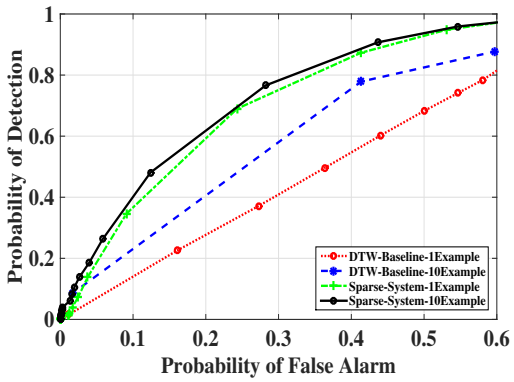

Figure 2: *ROC curves for the proposed sparse subspace detection system and the baseline DTW system [5] using different numbers of query examples available.*

5. Conclusions and Future Work

We proposed a novel QbE-STD method based on sparse reconstruction for classification. In contrast to the state of the art template matching methods, we cast the problem as subspace detection where the query and background subspaces are modeled through dictionary learning. Sparse representation of every frame of test posterior exemplars using the dictionaries characterizing the space of query and background enables discrimination of the underlying subspaces, and frame-level classification is achieved based on sparse reconstruction errors. Query decision is then simply performed by accumulating frame-level decisions (frames belonging to the query) over the hypothesized template, which results in a simple decoding process without any HMM. In spite of this simplicity (and potential for improvements), the proposed approach outperforms 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 QbE-STD tasks for international benchmarking [28]. To that end, discriminative dictionary learning techniques can be considered to reduce the ambiguities of the shared phonetic subspaces. We will also study integration of the phone posterior exemplars with sub-phonetic features, such as phonological posteriors, to address the adverse acoustic variability of multilingual non-native speech. Furthermore, we plan to devise alternative decision making procedure to exploit the temporal dependency of adjacent frames.

6. 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.

7. References

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

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

[3] I. Szoke, M. Fapso, L. 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*, 2008, pp. 42–48.

[4] X. Anguera, L. J. Rodriguez-Fuentes, I. Szoke, A. Buzo, F. Metze, and M. 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] L. J. Rodriguez-Fuentes, A. Varona, M. Penagarikano, G. Bordel, and M. 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] L. R. Rabiner, A. E. Rosenberg, and S. 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] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams," in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2009, pp. 398–403.

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

[9] C.-a. Chan and L.-s. 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] C.-y. Lee and J. 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] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210–227, 2009.

[12] Y. Chen, N. M. Nasrabadi, and T. 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] T. N. Sainath, B. Ramabhadran, M. Picheny, D. Nahamoo, and D. 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] J. F. Gemmeke, T. Virtanen, and A. 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] A. Asaei, "Model-based sparse component analysis for multiparty distant speech recognition," Ph.D. dissertation, École Polytechnique Fédéral de Lausanne (EPFL), 2013.

[16] D. Ram, A. Asaei, P. Dighe, and H. Bourlard, "Sparse modeling of posterior exemplars for keyword detection," in *Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015.

[17] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The ami meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005.

[18] P. Dighe, G. Luyet, A. Asaei, and H. Bourlard, "Exploiting low-dimensional structures to enhance dnn based acoustic modeling in speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2016.

[19] G. Luyet, P. Dighe, A. Asaei, and H. Bourlard, "Low-rank representation of nearest neighbor phone posterior probabilities to enhance DNN acoustic modeling," Tech. Rep. 217546, 2016, [online] https://infoscience.epfl.ch/record/217546.

[20] E. Elhamifar and R. 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] P. Dighe, A. Asaei, and H. Bourlard, "Sparse modeling of neural network posterior probabilities for exemplar-based speech recognition," *Speech Communication*, 2015.

[22] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research (JMLR)*, vol. 11, pp. 19–60, 2010.

[23] H. Ketabdar, J. Vepa, S. Bengio, and H. Bourlard, "Posterior based keyword spotting with a priori thresholds," in *International Conference on Spoken Language Processing (ICSLP)*, 2006.

[24] "Ami corpus partition," http://groups.inf.ed.ac.uk/ami/corpus/datasets.shtml.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding (ASRU)*, 2011.

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

[27] D. Ram, A. Asaei, and H. Bourlard, "Sparse subspace modeling for query by example spoken term detection," Idiap, Idiap-RR Idiap-RR-01-2016, 1 2016.

[28] M. A. Larson, B. Ionescu, M. Sjöberg, X. Anguera, J. Poignant, M. Riegler, M. Eskevich, C. Hauff, R. F. E. Sutcliffe, G. J. F. Jones, Y. Yang, M. Soleymani, and S. Papadopoulos, Eds., *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15, 2015*, ser. CEUR Workshop Proceedings, vol. 1436. CEUR-WS.org, 2015. [Online]. Available: http://ceur-ws.org/Vol-1436