SEMI-SUPERVISED TERM-WEIGHTED VALUE RESCORING FOR KEYWORD SEARCH

Kartik Audhkhasi[1], *Abhinav Sethy*[2], *Bhuvana Ramabhadran*[2], *Shrikanth S. Narayanan*[1]

[1]Signal Analysis and Interpretation Lab (SAIL)
Electrical Engineering Department
University of Southern California, Los Angeles, CA

[2]IBM T. J. Watson Research Center
Yorktown Heights, New York, NY

ABSTRACT

We present a semi-supervised algorithm for rescoring the output of a speech keyword search (KWS) system. Conventional loss functions such as squared-error and logistic loss are not suitable for optimizing the commonly-used KWS term-weighted value (TWV) performance metric. We derive a novel concave modified logistic log-likelihood function which lower-bounds TWV. We then use a manifold-regularized kernel classifier that maximizes this lower-bound. A manifold regularization term in our objective function uses available unlabeled speech data and makes our approach semi-supervised. This term is particularly useful for KWS in low-resource languages and ensures that the predicted keyword confidence scores are smooth on a low-dimensional manifold in the feature space. We conduct KWS experiments on the IARPA Babel Vietnamese task and show performance improvements in terms of the maximum TWV (MTWV). Our estimated confidence score is complementary with respect to the ASR posterior score and gives MTWV improvement upon interpolation with it.

Index Terms— Keyword search, term-weighted value, kernel methods, manifold regularization, semi-supervised learning.

1. INTRODUCTION

Keyword search (KWS) from speech is an information retrieval task that involves finding all possible locations of a given query term in a large speech data set. State-of-the-art KWS systems first decode the speech data set into word lattices using an automatic speech recognition (ASR) system. These word lattices are then converted into a finite state transducer (FST) index [1, 2] that stores the temporal locations and ASR posterior scores of all possible sequences of words (factors) in the lattices. The test-time input query term is then composed with this FST index to generate all putative keyword hits with their ASR posterior scores and time locations.

The quality of the ASR posterior score is often variable and depends on several factors. For example, out-of-vocabulary (OOV) queries are more likely to have unreliable posterior scores as compared to in-vocabulary (IV) queries because they are not part of the ASR system's vocabulary. Several works have focused on detecting and dealing with OOVs in KWS systems [3–6]. Longer queries post similar problems to the KWS system. Hence, the focus of this paper is on finding a complementary confidence score for KWS by *rescoring* the set of hits generated by the KWS system.

Our contributions in this paper focus on two salient characteristics of modern KWS systems. First, large-scale competitive evaluations such as STD-2006 [7] from NIST and the recent IARPA Babel program [8] use the term-weighted value (TWV) as a performance metric for KWS. Consider a query list $\mathcal{Q}$ and a detection threshold θ on the confidence scores of the retrieved hits. Then the TWV at θ is

$$TWV(\theta) = 1 - \frac{1}{|\mathcal{Q}|} \sum_{\mathcal{T} \in \mathcal{Q}} \Big(P_{\text{Miss}}(\mathcal{T}, \theta) + \beta P_{\text{FA}}(\mathcal{T}, \theta) \Big) \quad (1)$$

where $\beta = 999.9$, and the miss and false alarm probabilities for a query term $\mathcal{T}$ are

$$P_{\text{Miss}}(\mathcal{T}, \theta) = 1 - \frac{N_{\text{Correct}}(\mathcal{T}, \theta)}{N_{\text{Ref}}(\mathcal{T})} \quad \text{and} \quad (2)$$

$$P_{\text{FA}}(\mathcal{T}, \theta) = \frac{N_{\text{Spurious}}(\mathcal{T}, \theta)}{T_{\text{Audio}} - N_{\text{Ref}}(\mathcal{T})} \; . \quad (3)$$

Here $N_{\text{Correct}}(\mathcal{T}, \theta)$, $N_{\text{Spurious}}(\mathcal{T}, \theta)$, and $N_{\text{Ref}}(\mathcal{T})$ are the number of correctly detected, spurious, and true occurrences of the term $\mathcal{T}$ in the T_{Audio} second long audio data set. TWV thus weights false alarms and misses unequally. It is also more sensitive to misses on important rare queries such as proper nouns. Previous works on rescoring KWS system outputs do not consider TWV in their learning framework. Norouzian et al. [9, 10] use a squared-error loss function and perform hit classification instead of predicting a confidence score. Tu et. al [11] use a support vector machine classifier with various acoustic and linguistic features for re-ranking the hits. Lee et. al [12, 13] don't use an explicit discriminative loss function for learning the confidence scores but perform a random walk over an acoustic similarity graph with hits as nodes.

The second key feature of modern KWS systems is the availability of large amounts of unlabeled speech data. This is especially the case for low-resource languages such as Vietnamese, where the amount of data labeled with correct/incorrect hits is significantly less compared to a resource-rich language such as English. Nourozian et. al [9] present a manifold-regularized kernel least squares classifier for dealing with this challenge. Their approach is motivated from the fact that the hit features lie on a low-dimensional manifold in an otherwise high-dimensional space. However, they use a least squares classifier and don't directly estimate hit confidence scores.

This paper makes the following contributions in view of the above two features of modern KWS systems:

The authors would like to thank Brian Kingsbury of the IBM T. J. Watson Research Center and Doğan Can of SAIL at the University of Southern California for their suggestions.

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

This effort uses the IARPA Babel Program Vietnamese language collection release babel107b-v0.7 full language pack.

- **TWV Lower-Bound:** We propose a novel concave logistic lower-bound to the TWV in Section 3.
- **Manifold-regularized TWV Logistic Regression:** We use the above TWV lower-bound to learn a manifold-regularized kernel logistic regression classifier for predicting hit confidence scores in $[0, 1]$.

The next section sets the mathematical notation and introduces the standard kernel logistic regression [14]. We then derive our logistic lower-bound to the TWV in Section 3. Section 4 describes the Babel Vietnamese data set, the KWS system used in our experiments, and our computation of the kernel function using acoustic features extracted from the audio. We discuss our KWS results in Section 5 and conclude the paper in Section 6.

2. BACKGROUND ON MANIFOLD-REGULARIZED KERNEL LOGISTIC REGRESSION

This section recasts the conventional kernel logistic regression formulation [14] to our KWS problem. Consider a given query $\mathcal{T}$ from a query list $\mathcal{Q}$. Let the KWS system return $N_{\text{Hit}}(\mathcal{T})$ hits in response to this query. Without loss of generality, let the first $l(\mathcal{T}) \leq N_{\text{Hit}}(\mathcal{T})$ hits be labeled with $y_i = 1$ or 0 for correct or incorrect hits respectively. The remaining $u(\mathcal{T}) = N_{\text{Hit}}(\mathcal{T}) - l(\mathcal{T})$ hits are unlabeled. In a typical KWS setting, queries belong to the training set have all labeled hits while those at test time have all unlabeled hits.

Let each hit i for the query $\mathcal{T}$ have an associated d-dimensional feature vector $\mathbf{x}_i$, such as acoustic-prosodic features from the audio corresponding to hit i. Let a non-negative definite, Hermitian symmetric (Mercer) kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ compute the similarity between two feature vectors. Then K gives rise to a reproducing kernel Hilbert space (RKHS) [15] $\mathcal{H}_K$. The task of our learning problem is to estimate a function $f : \mathbb{R} \rightarrow \mathbb{R}$ in the RKHS $\mathcal{H}_K$ that predicts the log-odds ratio of a given hit being correct. The logistic sigmoid transformation of $f(\mathbf{x}_i)$

$$\sigma(f(\mathbf{x}_i)) = [1 + e^{-f(\mathbf{x}_i)}]^{-1} \tag{4}$$

thus estimates the probability $P(\text{hit } i \text{ correct})$. Kernel logistic regression estimates this function f by maximizing the log-likelihood of the labeled hits over all queries in $\mathcal{Q}$

$$f^* = \arg \max_{f \in \mathcal{H}_K} \frac{1}{|Q|} \sum_{\mathcal{T} \in \mathcal{Q}} \sum_{i=1}^{l(\mathcal{T})} \Big\{ \frac{y_i}{l(\mathcal{T})} \log \Big(\sigma(f(\mathbf{x}_i)) \Big) + \frac{(1 - y_i)}{l(\mathcal{T})} \log \Big(1 - \sigma(f(\mathbf{x}_i)) \Big) \Big\} \tag{5}$$

where $y_i \in \{1, 0\}$ is the correct/incorrect label of the hit. We refer to the log-likelihood function in (5) as $\mathcal{L}(f)$. The above optimization problem does not use the unlabeled hits for estimating f^*. Belkin, Niyogi, and Sindhwani [16] have however proposed including a manifold regularization term over both labeled and unlabeled data in the objective function in (5). This term becomes

$$\mathcal{M}(f) = \frac{1}{|Q|} \sum_{\mathcal{T} \in \mathcal{Q}} \frac{1}{N_{\text{Hhit}}(\mathcal{T})^2} \sum_{i,j=1}^{N_{\text{Hit}}(\mathcal{T})} \Big(f(\mathbf{x}_i) - f(\mathbf{x}_j) \Big)^2 W(\mathbf{x}_i, \mathbf{x}_j) \tag{6}$$

for our KWS setting. It is motivated from manifold learning that assumes the feature vectors $\mathbf{x}_i$ to lie on a low-dimensional manifold embedded in $\mathbb{R}^d$. Each hit lies at a vertex of an undirected graph on this manifold with edge weights $W(\mathbf{x}_i, \mathbf{x}_j)$. In our KWS formulation, minimizing this manifold smoothness penalty forces hits with close feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ (i.e. with high edge weight $W(\mathbf{x}_i, \mathbf{x}_j)$) to have similar values of the confidence score f. The overall manifold-regularized optimization problem thus becomes

$$f^* = \arg \max_{f \in \mathcal{H}_K} \Big\{ \mathcal{L}(f) - \gamma_1 ||f||_K^2 - \gamma_2 \mathcal{M}(f) \Big\} \tag{7}$$

where $||f||_K$ is the norm of f induced by the ambient inner product in the RKHS, and γ_1,γ_2 are non-negative real numbers. The representer theorem [16] converts the optimization problem in (7) to a problem of finding the optimal linear combination of kernel functions in a $\Big[\sum_{\mathcal{T} \in \mathcal{Q}} N_{\text{Hit}}(\mathcal{T}) \Big]$-dimensional space.

Our key focus for the next section will be the logistic log-likelihood function $\mathcal{L}(f)$ in (5). We immediately note that it is very different from the TWV function in (1). We thus propose a novel lower-bound to the TWV function in the next section.

3. LOGISTIC LOWER-BOUND ON THE TWV FUNCTION

We begin our derivation by relating the TWV function with the logistic loss by re-writing the former. Consider a given query term $\mathcal{T}$ and threshold θ on the predicted confidence score. Then

$$N_{\text{Correct}}(\mathcal{T}, \theta) = \sum_{i=1}^{l(\mathcal{T})} y_i I(\sigma(f(\mathbf{x}_i)) \geq \theta) \quad \text{and} \tag{8}$$

$$N_{\text{Spurious}}(\mathcal{T}, \theta) = \sum_{i=1}^{l(\mathcal{T})} (1 - y_i) I(\sigma(f(\mathbf{x}_i)) \geq \theta)$$
$$= a(\mathcal{T}) - \sum_{i=1}^{l(\mathcal{T})} (1 - y_i) I(\sigma(f(\mathbf{x}_i)) < \theta) \tag{9}$$

where $a(\mathcal{T})$ is a term independent of f and $I(.)$ is the indicator function. The inequalities within the indicator function can be re-written in terms of f using (4):

$$\sigma(f(\mathbf{x}_i)) \geq \theta \iff f(\mathbf{x}_i) \geq \log \Big(\frac{\theta}{1 - \theta} \Big) = c \, . \tag{10}$$

We now lower-bound the non-differentiable, non-convex indicator function $I(f(\mathbf{x}_i) \geq c)$ using the following inequalities as depicted in Figure 1:

$$I(f(\mathbf{x}_i) \geq c) \geq \log(\sigma(f(\mathbf{x}_i) - c)) \quad \text{and} \tag{11}$$
$$I(f(\mathbf{x}_i) < c) \geq \log(1 - \sigma(f(\mathbf{x}_i) - c)) \, . \tag{12}$$

The above bounds give the following bounds on the number of correct and spurious hits of the term $\mathcal{T}$:

$$N_{\text{Correct}}(\mathcal{T}, \theta) \geq \sum_{i=1}^{l(\mathcal{T})} y_i \log(\sigma(f(\mathbf{x}_i) - c)) \quad \text{and} \tag{13}$$

$$N_{\text{Spurious}}(\mathcal{T}, \theta) \leq a(\mathcal{T}) - \sum_{i=1}^{l(\mathcal{T})} (1 - y_i) \log(1 - \sigma(f(\mathbf{x}_i) - c)) \, . \tag{14}$$

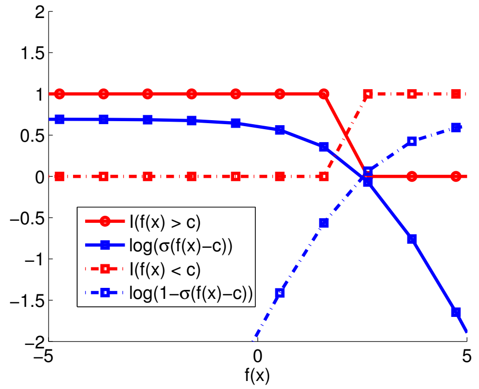

Fig. 1. This figure shows the indicator functions $I(f(\mathbf{x}) \geq c)$ and $I(f(\mathbf{x}) < c)$ with their logistic lower-bounds in (11) and (12) for $c = 2.5$. We have added $\log(2)$ to the bounds to ensure that they touch the indicator functions at $f(\mathbf{x}) = c$.

Hence the miss and false alarm probabilities are upper-bounded as:

$$P_{\text{Miss}}(\mathcal{T}, \theta) \leq 1 - \frac{1}{N_{\text{ref}}(\mathcal{T})} \sum_{i=1}^{l(\mathcal{T})} y_i \log(\sigma(f(\mathbf{x}_i) - c)) \quad \text{and} \tag{15}$$

$$P_{\text{Spurious}}(\mathcal{T}, \theta) \leq a(\mathcal{T}) - \frac{1}{T_{\text{Audio}} - N_{\text{Ref}}(\mathcal{T})} \sum_{i=1}^{l(\mathcal{T})} (1 - y_i) \times \log(1 - \sigma(f(\mathbf{x}_i) - c)) \, . \tag{16}$$

Substituting the above bounds in the TWV expression (1) gives us the following logistic lower-bound on TWV:

$$TWV(\mathcal{T}, \theta) \geq \frac{1}{|\mathcal{Q}|} \sum_{\mathcal{T} \in \mathcal{Q}} \sum_{i=1}^{l(\mathcal{T})} \left\{ \frac{y_i}{N_{\text{Ref}}(\mathcal{T})} \log\Big(\sigma(f(\mathbf{x}_i) - c)\Big) + \frac{\beta(1 - y_i)}{T_{\text{Audio}} - N_{\text{Ref}}(\mathcal{T})} \log\Big(1 - \sigma(f(\mathbf{x}_i) - c)\Big) - \beta a(\mathcal{T}) \right\} . \tag{17}$$

We can now compare the above lower-bound on TWV with the standard kernel logistic regression objective function in (5). There are two key differences. First, the TWV lower-bound weights the $y_i = 1$ and $y_i = 0$ terms unequally, as compared to the equal weight of $1/l(\mathcal{T})$ in (5). These unequal weights reflect the relative importance of misses and false alarms in the TWV function. Second, we observe that the detection threshold θ appears through c in (17) compared to the default $\theta = 0.5$ or $c = 0$ in the standard kernel logistic regression log-likelihood function.

Maximizing the TWV logistic lower-bound in (17) gives us the following kernel logistic regression optimization problem for TWV:

$$f^* = \arg\max_{f \in \mathcal{H}_K} \frac{1}{|\mathcal{Q}|} \sum_{\mathcal{T} \in \mathcal{Q}} \sum_{i=1}^{l(\mathcal{T})} \left\{ \frac{y_i}{N_{\text{Ref}}(\mathcal{T})} \log\Big(\sigma(f(\mathbf{x}_i) - c)\Big) + \frac{\beta(1 - y_i)}{T_{\text{Audio}} - N_{\text{Ref}}(\mathcal{T})} \log\Big(1 - \sigma(f(\mathbf{x}_i) - c)\Big) \right\} . \tag{18}$$

Inclusion of the manifold regularization term from (6) and the ambient norm of f with the representer theorem again gives a concave optimization problem in the kernel weights $\boldsymbol{\alpha}$. We note that $N_{\text{Ref}}(\mathcal{T})$ and T_{audio} are needed for queries corresponding to all labeled hits in the above objective function. However, this information is not needed at test time because the estimated confidence score depends only on the estimated $\boldsymbol{\alpha}^*$ and the feature vectors for all hits used during training. The next section describes the Babel Vietnamese data set, our KWS system, and computation of the kernel function between hits.

4. EXPERIMENTAL SETUP

4.1. Babel Vietnamese Data Set

Vietnamese was the surprise language of the Babel OpenKWS13 evaluation. The mono-syllabic and tonal nature of Vietnamese make it a challenging language for automated spoken language processing. Vietnamese also has many regional dialects with subtle differences. We focused on the Vietnamese full language pack (FullLP) for experiments in this paper. The FullLP contains 20 hours of word-transcribed scripted speech, 80 hours of word-transcribed conversational telephone speech, and a pronunciation lexicon. We used an automatically-generated a list of 200 keywords for this development data set as reported in [17]. We focused on the test audio reuse (TAR) scenario where the KWS system is allowed to re-process the test audio after the keyword hits are returned.

4.2. Speaker-Adapted Deep Neural Network ASR and FST-based KWS System System

We used a similar speaker-adapted (SA) deep neural network (DNN) ASR system as described in [18] for Cantonese. The system uses a baseline Gaussian mixture model-hidden Markov model (GMM-HMM) system trained using a standard pipeline in IBM's Attila toolkit [19]. The first DNN training stage minimizes the cross-entropy between the quinphone context-dependent HMM state targets and the output layer activations of the neural network using backpropagation [20]. The DNN model is finally trained with the state-level minimum Bayes risk (MBR) criterion with a distributed implementation [21] of Hessian-free optimization. We use the trained DNN with 3000 softmax output quinphone HMM states in a hybrid configuration [22].

We used a two-pass implementation [23] of the weighted FST audio indexing and search algorithm in [1]. The system uses two indices - a word index for in vocabulary (IV) queries and a phonetic index for OOV queries. The output of the KWS system is a *postings list* which contains a list of hits for each query keyword, the associated ASR posterior score, start time, and end time.

4.3. Kernel Function Computation

Our kernel machine framework in (7) and Section 3 does not require computation of a feature vector $\mathbf{x}_i$ for each hit i due to the representer theorem [16], but the kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ for all pairs (i, j) of hits. We used the generalized RBF kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\Big\{ - D(\mathbf{x}_i, \mathbf{x}_j)^2 / (2\sigma_K^2) \Big\} \tag{19}$$

where $D(\mathbf{x}_i, \mathbf{x}_j)$ is the dissimilarity between hits i and j. We note that D need not be a distance metric but K should be a valid Mercer's kernel. We address this issue later in this section.

We computed two hit dissimilarity measures for this work. The first measure uses the hit audio and is the normalized dynamic time warping (DTW) mean squared error between perceptual linear prediction (PLP) feature vector sequences for the two hits. We extracted

13-dimensional PLP coefficients over 25 msec frames with 10 msec shift from the audio using Kaldi's [24] feature extraction tool. We then aligned the PLP sequences from the two given hits using DTW and found the total squared-error cost of the best alignment path. We then normalized this cost by the length of the diagonal of the cost matrix and divided the resulting score by the feature dimension 13. Let us represent the resulting dissimilarity score between hits i and j as $D_{\text{DTW}}(\mathbf{x}_i, \mathbf{x}_j)^2$.

The second hit dissimilarity measure assigns higher similarity to hits belonging to the same query as compared to different queries. We define

$$D_{\text{Clique}}(i,j)^2 = \begin{cases} 0 & \text{; if i and j belong to the same query} \\ \zeta & \text{; otherwise} \end{cases} \tag{20}$$

where $\zeta > 0$ is a tunable parameter. Setting $\zeta \rightarrow \infty$ assigns infinite dissimilarity to hits belonging to different queries. Hence the kernel optimization problem splits into many disjoint optimization problems over each query term. Setting $\zeta = 0$ removes any distinction between hits belonging to different queries in the learning algorithm.

We compute the kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ using the above two hit dissimilarity functions as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{ \frac{-[D_{\text{DTW}}(\mathbf{x}_i, \mathbf{x}_j)^2 + D_{\text{Clique}}(i,j)^2]}{2\sigma_K^2} \right\}. \tag{21}$$

While an RBF kernel function using squared-Euclidean distance between vectors is non-negative definite, the above kernel using DTW dissimilarity is not guaranteed to be so. Hence we make this kernel matrix non-negative definite by performing its eigenvalue decomposition, setting all negative eigenvalues to a small positive constant, and reconstructing the kernel matrix. We observed only a marginal difference between the original and reconstructed kernel matrices because most of the large eigenvalues were positive. We next describe our KWS results and related analysis in the next section.

5. RESULTS AND DISCUSSION

We split the postings list generated by the KWS system into three disjoint subsets of queries for training, testing, and development. We performed 3-fold cross-validation by labeling hits from one set as correct/incorrect (the training set), and using the other two sets without any labels (the testing and development sets) in our learning framework. We fixed the RBF kernel standard deviations for the graph weight matrix $\mathbf{W}$ and kernel matrix $\mathbf{K}$ to 0.5. We varied the weights of the two regularization terms (γ_1 and γ_2) over $\{0, 1e^{-5}, 1e^{-4}, 1e^{-3}, 1e^{-2}, 1e^{-1}, 1\}$ and clique dissimilarity parameter ζ over $\{0, 1, 2\}$. We picked the best-performing hyperparameter values using the development set loss function value (excluding the two regularization terms) in the learning objective function. We did not use conventional classification performance metrics such as the F1 score or equal error rate to do this selection because they do not measure the loss function of interest.

Table 1 lists the maximum TWV (MTWV) of the entire posting list re-scored with manifold-regularized kernel logistic regression using the standard logistic regression loss (MR-LL-KLR) and the proposed TWV approximation (MR-TWV-KLR). We used the popular sum-to-one (STO) normalization [17] of the posting list before evaluation. We observe a 4.8% increase in MTWV after using the TWV loss.

We then further analyzed the benefit of the posting list re-scored by MR-TWV-KLR by evaluating the MTWV after interpolation with the ASR posterior scores. We implemented a simple mixture-of-experts (MOE) interpolation scheme where the output score of a hit i for a term $\mathcal{T}$ is

Rescoring System	MTWV
Manifold-regularized logistic loss kernel logistic regression (MR-LL-KLR)	0.2788
Manifold-regularized TWV loss kernel logistic regression (MR-TWV-KLR)	**0.2913** **(+4.8%)**

Table 1. This table shows the MTWV of the manifold-regularized kernel logistic regression system using the standard logistic loss and the proposed TWV loss on the Vietnamese development data set.

$N_0(\mathcal{T})$	MTWV
0	0.3551 (+1.0%)
2	0.3567 (+1.4%)
4	**0.3578 (+1.8%)**
6	0.3550 (+1.0%)
8	0.3521 (+0.1%)
10	0.3516 (+0%)
∞ (ASR posterior)	0.3516

Table 2. This table shows the MTWV of the best manifold-regularized kernel logistic regression system using the proposed TWV loss (MR-TWV-KLR) after mixture-of-experts interpolation with the ASR posterior scores using (22) for different values of parameter $N_0(\mathcal{T})$.

$$s_i(\mathcal{T}) = \alpha\, \sigma(f^*(\mathbf{x}_i) - c^*) + (1-\alpha)\, p(i) \tag{22}$$

$$\text{where} \quad \alpha = 0.1\sigma(N_{\text{Hits}}(\mathcal{T}) - N_0(\mathcal{T}))\ , \tag{23}$$

$p(i)$ is the ASR posterior score of hit i, and $N_0(\mathcal{T})$ is the number of hits at which the MR-TWV-KLR score gets weight $\alpha = 0.05$. This MOE fusion rule gives higher weight to the MR-TWV-KLR score for queries with more hits. Higher values of $N_0(\mathcal{T})$ cause the MR-TWV-KLR score to be emphasized only for queries with high number of hits. Table 2 shows that the optimal $N_0(\mathcal{T})$ is 4 hits. Reducing $N_0(\mathcal{T})$ reduces performance because the MR-TWV-KLR score is being given more emphasis than is optimal for queries with just $1-3$ hits. Our estimate score is unreliable for such queries because of small query-clique size in the hit similarity graph.

6. CONCLUSION AND FUTURE WORK

We presented a novel lower-bound to the popular TWV performance metric for modern KWS systems. We then maximized this lower-bound in a semi-supervised kernel method for rescoring a KWS system. This in contrast to prior work which does not explicitly optimize the TWV. The proposed algorithm is semi-supervised because it uses unlabeled data through a manifold regularization term in the objective function. Our algorithm gives a 4.8% improvement in MTWV over a manifold-regularized kernel logistic regression baseline. The generated confidence scores are complementary to the ASR posterior scores and further improve their MTWV by 1.8%.

Future work should focus on deriving better lower-bounds to the TWV, incorporating the ASR posterior scores in the kernel learning framework, designing better hit dissimilarity measures, and extending this framework to the fusion of multiple KWS systems.

7. REFERENCES

[1] C. Allauzen, M. Mohri, and M. Saraclar, "General indexation of weighted automata: application to spoken utterance retrieval," in *Proc. Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL*. ACL, 2004, pp. 33–40.

[2] D. Can and M. Saraclar, "Lattice indexing for spoken term detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2338–2347, 2011.

[3] J. Mamou, B. Ramabhadran, and O. Siohan, "Vocabulary independent spoken term detection," in *Proc. SIGIR*. ACM, 2007, pp. 615–622.

[4] C. Parada, A. Sethy, M. Dredze, and F. Jelinek, "A spoken term detection framework for recovering out-of-vocabulary words using the web.," in *Interspeech*, 2010, pp. 1269–1272.

[5] C. Parada, A. Sethy, and B. Ramabhadran, "Query-by-example spoken term detection for OOV terms," in *Proc. ASRU*. IEEE, 2009, pp. 404–409.

[6] C. Parada, M. Dredze, D. Filimonov, and F. Jelinek, "Contextual information improves OOV detection in speech," in *Proc. HLT-NAACL*. ACL, 2010, pp. 216–224.

[7] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddingtion, "Results of the 2006 spoken term detection evaluation," in *Proc. ACM SIGIR Workshop on Searching Spontaneous Conversational*, 2007, pp. 51–55.

[8] "IARPA - Babel Program," `http://www.iarpa.gov/Programs/ia/Babel/babel.html`.

[9] A. Norouzian, R. Rose, and A. Jansen, "Semi-supervised manifold learning approaches for spoken term verification," in *Proc. Interspeech*, 2013, pp. 2594–2598.

[10] A. Norouzian, R. Rose, S. H. Ghalehjegh, and A. Jansen, "Zero resource graph-based confidence estimation for open vocabulary spoken term detection," in *Proc. ICASSP*, 2013.

[11] T. Tu, H. Lee, and L. Lee, "Improved spoken term detection using support vector machines with acoustic and context features from pseudo-relevance feedback," in *Proc. ASRU*, 2011, pp. 383–388.

[12] H. Lee and L. Lee, "Improved semantic retrieval of spoken content by document/query expansion with random walk over acoustic similarity graphs," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2013.

[13] H. Lee, Y. Chen, and L. Lee, "Improved speech summarization and spoken term detection with graphical analysis of utterance similarities," in *Proc. APSIPA ASC*.

[14] J. Zhu and T. Hastie, "Kernel logistic regression and the import vector machine," in *Proc. NIPS*, 2001, pp. 1081–1088.

[15] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, no. 3, pp. 337–404, 1950.

[16] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *The Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[17] J. Mamou, J. Cui, X. Cui, M. J. F. Gales, B. Kingsbury, K. Knill, L. Mangu, D. Nolden, M. Picheny, B. Ramabhadran, M. Saraclar, R. Schluter, A. Sethy, and P. C. Woodland, "Developing keyword search under the IARPA Babel program," in *Proc. Afeka Speech Processing Conference*, 2013.

[18] J. Cui, X. Cui, J. Mamou, B. Kingsbury, B. Ramabhadran, L. Mangu, M. Picheny, A. Sethy, and J. Kim, "Developing speech recognition systems for corpus indexing under the IARPA Babel program," in *Proc. ICASSP*, 2013.

[19] H. Soltau, G. Saon, and B. Kingsbury, "The IBM Attila speech recognition toolkit," in *Proc. SLT*, 2010, pp. 97–102.

[20] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. ASRU*, 2011, pp. 24–29.

[21] B. Kingsbury, T. N. Sainath, and H. Soltau, "Scalable minimum Bayes risk training of deep neural network acoustic models using distributed hessian-free optimization.," in *Proc. Interspeech*, 2012.

[22] T. N. Sainath, B. Kingsbury, B. Ramabhadran, P. Fousek, P. Novak, and A.-R. Mohamed, "Making deep belief networks effective for large vocabulary continuous speech recognition," in *Proc. ASRU*, 2011, pp. 30–35.

[23] D. Can, E. Cooper, A. Sethy, C. White, B. Ramabhadran, and M. Saraclar, "Effect of pronounciations on OOV queries in spoken term detection," in *Proc. ICASSP*, 2009, pp. 3957–3960.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz, "The Kaldi Speech Recognition Toolkit," in *Proc. ASRU*. Dec. 2011, IEEE.