TOWARD UNSUPERVISED MODEL-BASED SPOKEN TERM DETECTION WITH SPOKEN QUERIES WITHOUT ANNOTATED DATA

Chun-an Chan[⋆] *Cheng-Tao Chung*[†] *Yu-Hsin Kuo*[#] *Lin-shan Lee*[⋆†#]

[⋆] Graduate Institute of Communication Engineering, National Taiwan University
[†] Graduate Institute of Electrical Engineering, National Taiwan University
[#] Department of Electrical Engineering, National Taiwan University

chunanchan@gmail.com, r01921031@ntu.edu.tw, pandakuogo@gmail.com, lslee@gate.sinica.edu.tw

ABSTRACT

We present a two-stage model-based approach for unsupervised query-by-example spoken term detection (STD) without any annotated data. Compared to the prevailing DTW approaches for the unsupervised STD task, HMMs used by model-based approaches can better capture the signal distributions and time trajectories of speech with a more global view of the spoken archive; matching with model states also significantly reduces the computational load. The utterances in the spoken archive are first offline decoded into acoustic patterns automatically discovered in an unsupervised way from the spoken archive. In the first stage, we propose a document state matching (DSM) approach, where query frames are matched to the HMM state sequences for the spoken documents. In this process, a novel duration-constrained Viterbi (DC-Vite) algorithm is proposed to avoid unrealistic speaking rate distortion. In the second stage, pseudo relevant/irrelevant examples retrieved from the first stage are respectively used to construct query/anti-query HMMs. Each spoken term hypothesis is then rescored with the likelihood ratio to these two HMMs. Experimental results show an absolute 11.8% of mean average precision improvement with a more than 50% reduction in computation time compared to the segmental DTW approach on a Mandarin broadcast news corpus.

Index Terms— Unsupervised spoken term detection, zero-resource, query-by-example, speech pattern discovery

1. INTRODUCTION

The fast growing quantity of video and audio content over the Internet demands efficient and accurate approaches to search through the spoken contents. The spoken term detection (STD) task is to find all occurrences of the orthographically specified query terms from a large spoken archive [1]. Most STD systems were based on automatic speech recognition (ASR), transforming speech into words or subwords for token matching [2, 3, 4, 5, 6]. The performance of these methods relies heavily on the performance of the speech recognizer [7], thus requiring large orthographically transcribed training data. Recently, there have been efforts in query-by-example (QbE) STD, where a spoken query is provided instead of a text query [8, 9]. However this further complicates the problem because short queries are error-prone for ASR and usually contain out-of-vocabulary words. These methods are not reliable especially for those languages with very limited annotated data [10] or some dialects that have no writing systems [11].

Considering the above difficulties, there have been recent efforts in QbE STD without speech recognition [12, 13], which is also the focus of this work. Hereafter we assume all queries are in speech form, and no annotated speech data is available. Prevailing approaches to this task rely on dynamic time warping (DTW) to directly match the spoken query to the spoken documents based on the signal characteristics. This removes the requirement for annotated data and the impact of recognition errors. However, a major limitation of DTW is that the distances are easily affected by speaker mismatch and acoustic conditions. Many related works focused on feature representations and distance measures that are more invariant to speaker and acoustic condition diversities within the DTW framework [14], including the posteriorgrams of a universal Gaussian mixture model [15], and the acoustic segment models [16, 17]. These methods employed DTW in their matching processes, which essentially take computation time linear to the number of frames to be searched. Substantial efforts were devoted to reducing the computation load, including a segment-based DTW [18, 19], a lower-bound estimation for DTW [20, 21], and a locality sensitive hashing technique for indexing speech frames [22].

In this work, we solve the unsupervised QbE STD problem for the first time without DTW, using a set of models generated by automatic acoustic pattern discovery, including a subword-like pattern acoustic model, a word-like pattern lexicon, and a word-like pattern N-gram model [23]. Compared to DTW, the HMM is well known for its ability to model the distributions and time trajectories of speech signals, which better handles the signal variation problem in DTW. The utterances in the spoken archive are decoded offline into word-like patterns, followed by two online stages. In the first stage, we propose a **document state matching** (DSM) approach to match query frames to the decoded states in the documents. In this way, not only do the HMMs better model the signal distributions and time trajectories, the much smaller number of states than frames for the documents leads to much lower computational load. Although the above DSM can be realized with the Viterbi algorithm, we propose a **duration-constrained Viterbi** (DC-Vite) algorithm to avoid unrealistic speaking rate distortion between the query and the detected spoken term. In the second stage, we propose a **pseudo likelihood ratio** (PLR) approach by evaluating for each hypothesized spoken term the likelihood ratio to the query/anti-query HMMs trained with the pseudo relevant/irrelevant examples obtained in the first stage. Significant detection improvements were observed with much less computation time compared to the segmental DTW approach [15]. The results highlight the potential of leveraging well-developed HMM-based speech processing techniques for model-based approaches for zero-resource STD. The detection performance also suggests the feasibility of indexing speech data with acoustic patterns.

2. DOCUMENT STATE MATCHING

The aim of document state matching (DSM) is to use HMM state representations of the spoken documents that have a more global view over the speech signals in the archive than the relatively local view of frame-based representation used in DTW-based approaches. The models used to decode the spoken archive in the preprocessing stage consist of a set of subword-like pattern HMMs, a lexicon whose word-like patterns are expanded as subword-like patterns, and a word-like pattern N-gram model, all generated from the spoken archive by an acoustic pattern discovery approach [23]. The documents in the archive are first offline decoded into HMM state sequences of subword-like patterns. When the user enters a spoken query, we find a partial state sequence in the document that maximizes the likelihood of generating the query.

2.1. Viterbi Decoding on Document State Sequence

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathbf{X}|}\}$ be the query frame sequence and $P = (q_1, \ldots, q_{|P|})$ be the document state sequence. The matching relation between the document states and the query frames is defined by a state assignment $Z = (j_1, j_2, \ldots, j_{|\mathbf{X}|})$ for $\mathbf{X}$ such that the query frame $\mathbf{x}_i$ is matched to the document state q_{j_i}. Our goal is to find a state assignment Z for $\mathbf{X}$ that maximizes the likelihood of the state assignment and query features,

$$l(\mathbf{X}, Z) = \Pr(q_{j_1}) \cdot \prod_{i=1}^{|\mathbf{X}|-1} \Pr(q_{j_{i+1}}|q_{j_i}) \prod_{i=1}^{|\mathbf{X}|} \Pr(\mathbf{x}_i|q_{j_i}). \quad (1)$$

The state assignment Z must be monotonic and can begin and terminate at any state, which corresponds to matching the query to a partial state sequence in P. This problem can be solved by the Viterbi algorithm on the index plane of document states (vertical axis) and query frames (horizontal axis) as in Fig. 1. A state assignment corresponds to a monotonic path going from the leftmost column to the rightmost column on this index plane. An example path going from $(1, 2)$ to $(15, 6)$ in Fig. 1 represents the matching of $(\mathbf{x}_1, \ldots, \mathbf{x}_{15})$ to $(q_2, \ldots, q_6)$.

2.2. Duration-Constrained Viterbi

Note in (1) there is no constraint on the number of frames matched to the same state. When applying Viterbi decoding on the above problem, the frame duration in each document state is modeled by an exponential probability distribution due to the self-transition probability, with known weakness [24]. In stead of the exponential duration modeling in the HMM, here we explicitly constrain the possible duration of each state based on the corresponding number of frames in the document being considered, referred to as duration-constrained Viterbi (DC-Vite).

Let f_j be the number of document frames matched to q_j in P. We require that the number of query frames matched to state q_j must be between $\frac{f_j}{\gamma}$ and γf_j, where γ controls the maximum possible speaking rate ratio. In the example of Fig. 1, $|P| = 6$ and $f_1 = 3$ (3 document frames correspond to q_1), $f_2 = 5$, $f_3 = 4$, etc. A horizontal line on the path represents a document state matched to multiple query frames. When computing the optimal partial paths ending at $(i, j) = (12, 3)$ with $f_3 = 4$ and $\gamma = 2$ in dynamic programming, the number of frames staying in q_3 must be between $\frac{f_3}{\gamma} = 2$ and $\gamma f_3 = 8$. Hence we examine all paths from $(i, j) = (4, 2), (5, 2), \ldots, (9, 2), (10, 2)$ and pick the one subpath that yields maximum likelihood when concatenated with the last horizontal path. In this way, the numbers of query and document frames matched to the same state differ at most by a factor of γ. Thus those paths corresponding to unrealistic speaking rate distortion are eliminated.

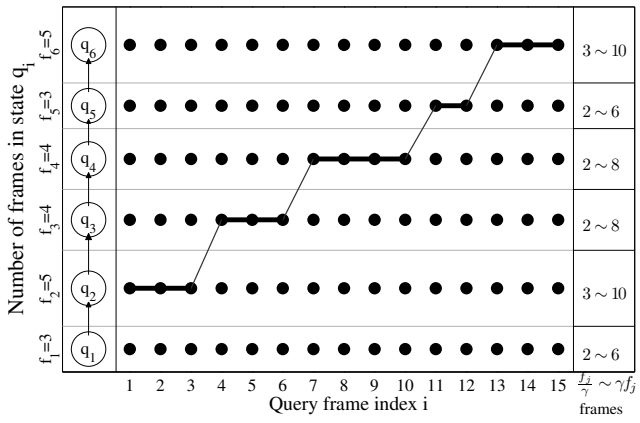

Fig. 1. An example of the index plane for query frames and a document state sequence with $\gamma = 2$. Each bold line represents a document state matched to several query frames. For each state q_j, the number of document frames f_j and the number of allowed query frames ($\frac{f_j}{\gamma} \sim \gamma f_j$) are respectively listed on the left and right ends. For example, state q_3 corresponds to 4 frames in the document ($f_3 = 4$), so it can match $2 \sim 8$ frames in the query given $\gamma = 2$.

To detect multiple matched regions in a document, we collect all optimal state assignments ending at states q_1 to $q_{|P|}$ and examine where they start. From each identical starting state, we choose one best state assignment with maximum likelihood $l(\mathbf{X}, Z)$. In this way, multiple spoken term regions appearing in the same utterance can be individually detected. The detected spoken term regions in the documents are referred to as **hypothesized regions** in this paper.

Aside from the fact that the proposed DSM approach represents documents with HMM state sequences trained with a more global view, this approach also provides an additional benefit in reducing computation. Both DTW and DC-Vite algorithms need to perform local distance/likelihood calculations: frame-wise distance for DTW, and query frame likelihood, $\Pr(\mathbf{x}_i|q_j)$, for DC-Vite. For DTW-based approaches, the number of computations for local distance required for a query is approximately the number of query frames times the number of frames in all documents. In contrast, the number of computations for $\Pr(\mathbf{x}_i|q_j)$ required for a query is equal to the number of query frames times the number of Gaussians in all HMMs. This is due to the state representation for documents; all frame level information in the documents is encoded into the states. Also the computations for likelihood will not grow linearly with the number of document frames, since $\Pr(x_i|q_j)$ is precomputed and stored once the query is entered, and reused for different documents. This is especially important for searching large spoken document archive. Such a scheme is also suitable for indexing a spoken archive for even faster document retrieval.

3. PSEUDO LIKELIHOOD RATIO

Following the DSM, a list of possible hypothesized regions (parts of signals hypothesized as the query term) are obtained, whose scores

describe their similarity to the query. These hypothesized regions are sorted by the scores, with more similar regions listed higher. A set $\mathcal{R}$ of **pseudo relevant examples** is defined by selecting the hypothesized regions in the list whose rankings (or scores) are higher than a threshold. Another set $\mathcal{I}$ of **pseudo irrelevant examples** are also defined by selecting hypothesized regions whose rankings (or scores) lie below a threshold. We assume these pseudo irrelevant examples are in fact not the desired spoken term, but are similar to the query in the signal space. We train two left-to-right HMMs, $\Lambda_{\mathcal{R}}, \Lambda_{\mathcal{I}}$ using the sets $\mathcal{R}$ and $\mathcal{I}$ for each query respectively. The model $\Lambda_{\mathcal{R}}$ is referred to as the query HMM and $\Lambda_{\mathcal{I}}$ as the anti-query model. After EM training of $\Lambda_{\mathcal{R}}$ and $\Lambda_{\mathcal{I}}$, each hypothesized region is rescored by the likelihood ratio to $\Lambda_{\mathcal{R}}$ and $\Lambda_{\mathcal{I}}$ [25].

We apply the same initialization steps for $\Lambda_{\mathcal{R}}$ and $\Lambda_{\mathcal{I}}$. They differ after EM training with different training sets $\mathcal{R}$ and $\mathcal{I}$. Following the notations from Subsection 2.2, assume the partial state sequence for the most similar pseudo relevant example is $Q = (q_1, q_2, \ldots, q_{|Q|})$, where q_k is the k-th state in the sequence. The query HMM $\Lambda_{\mathcal{R}}$ and the anti-query HMM $\Lambda_{\mathcal{I}}$, both with $|Q|$ states, are then initialized as follows. Let the states in $\Lambda_{\mathcal{R}}$ be $\{\widehat{q}_1, \ldots, \widehat{q}_{|Q|}\}$. The Gaussians in state $\widehat{q}_j$ are copied from the Gaussians in q_j without sharing. The prior probabilities are set as $\Pr(\widehat{q}_1) = 1$ and $\Pr(\widehat{q}_j) = 0$ otherwise. The self-transition probability of state $\widehat{q}_j$ is copied from state q_j, and the rest is set to the transition probability to the next state. Therefore $\Pr(\widehat{q}_{|Q|}|\widehat{q}_{|Q|}) = 1$, and for $j = 1, \ldots, |Q| - 1$,

$$\Pr(\widehat{q}_j|\widehat{q}_j) = \Pr(q_j|q_j) \quad (2)$$

$$\Pr(\widehat{q}_{j+1}|q_j) = 1 - \Pr(q_j|q_j) \quad (3)$$

The same initialization is applied on $\Lambda_{\mathcal{I}}$, so all the parameters of $\Lambda_{\mathcal{I}}$ have the same initial value as those in $\Lambda_{\mathcal{R}}$ before EM training.

After the model $\Lambda_{\mathcal{R}}$ is initialized, it is trained using the query feature sequence $\mathbf{X}_0$ and the pseudo relevant examples $\mathcal{R} = \{\mathbf{X}_1, \ldots, \mathbf{X}_{|\mathcal{R}|}\}$, where $\mathbf{X}_1$ has the highest score. Since the hypothesized regions have different confidence of being relevant to the query, the i-th training example $\mathbf{X}_i$ has a training weight $2^{-\lambda i}$, where λ controls the decay rate of weight based on the ranking. For anti-query model $\Lambda_{\mathcal{I}}$ training, the training weights for each training sample in $\mathcal{I}$ are set to be equal, so pseudo irrelevant examples are modeled with equal importance. Then EM algorithm is performed on both models.

After the two models are trained, the relevance score of each hypothesized region is set to the log likelihood ratio for the region with respect to the query HMM $\Lambda_{\mathcal{R}}$ and anti-query HMM $\Lambda_{\mathcal{I}}$. In this way, the finer signal differences that separate the true spoken terms from false accepted regions can contribute to larger likelihood ratios for the signal regions, while the likelihoods of signal parts that are inseparable among true and false hypotheses will be depressed.

4. EXPERIMENTAL SETUP

4.1. Queries and Document Archive

We evaluated the proposed approach with a spoken term detection task on Mandarin broadcast news. The audio archive was collected daily from the *News98 FM radio* station in Taiwan in August, 2001 and is 4.1 hours long in total. The news stories were manually segmented into 5034 utterances, which were taken as 5034 documents to be searched through. Except for a small number of utterances produced by male reporters, the spoken documents were all produced by female speakers. From the spoken documents, 42 test query instances were manually extracted, containing names of countries, events, politicians and organizations. The test query terms ranged in length from 2 to 7 syllables, with a majority of 2 and 3 syllables. The number of relevant documents for each query ranged from 9 (0.2%) to 111 (2.2%) , averaging 26 (0.5%) in the archive. When evaluating the detection performance of each query, the document containing the query instance was excluded from searching and evaluation. To obtain the necessary parameters in the approaches discussed here, ten additional queries were defined in the Mandarin broadcast news corpus as the development set. All parameters used in our experiments were obtained using the development set.

Table 1. The parameters for the model generated by two-level acoustic pattern discovery.

model / parameter	count
#word-like patterns	362
#subword-like pattern models	208
#state per subword-like pattern model	13
#total states (include sil & sp)	2707
#total Gaussians (include sil & sp)	2713
language model	bigram

4.2. Feature and Evaluation Metric

In the baseline segmental DTW experiment, we used Gaussian posteriorgrams as the feature vectors and the frame-wise distance measure $d(\mathbf{x}, \mathbf{y}) \equiv -\log(\mathbf{x} \cdot \mathbf{y})$, as suggested by previous works [14]. The Gaussian posteriorgrams were obtained from a universal Gaussian mixture model (GMM). The feature vectors used to train the GMM were the conventional 13-dimensional MFCC features with a 20ms Hamming window and a 10ms frame shift, concatenated with the derivative and acceleration. Fifty Gaussians were used in the GMM [15]. The same feature vectors were used in training subword-like pattern HMMs.

The mean average precision (MAP) [26] was used to evaluate the detection performance. To calculate the MAP value, all documents were ranked according to their relevance to the query. For DTW approaches, lower distance scores imply higher relevance, whereas for our model-based approach, higher likelihoods or likelihood ratios imply higher relevance. Hits and misses were evaluated per document. That is, a hit was counted if the returned document contained the desired query term, regardless of whether it contained single or multiple spoken terms. The MAP values reported in this paper were obtained using the `trec_eval` toolkit[1]. We also evaluate the CPU time for our proposed approaches. The unit of CPU time is sec/DHQS (document-hour $\times$ query-second), that is, the computation time in seconds if an hour of document is search using a one second long query.

4.3. Model of Acoustic Patterns

The parameters of the model generated by acoustic pattern discovery [23] are listed in Table 1. There are 208 subword-like patterns, which are highly correlated to Mandarin syllables. The HMM for each subword-like model has 13 states, each with single Gaussian component. There are 362 word-like patterns in the lexicon; the 154 extra word-like patterns are composed of multiple subword-like patterns.

[1] [Online]. Available: http://trec.nist.gov/trec_eval/

Table 2. MAP(%) and average CPU time of DTW-based and proposed model-based approach. Row (1) are the results for baseline segmental DTW approaches. Rows (2)-(3) are the results using proposed DSM. Row (4)-(5) show the results of PLR by rescoring or score fusion. The unit of CPU time is sec/DHQS (document-hour $\times$ query-second).

	Method	MAP	P@10	CPU time
(1)	Segmental DTW	37.5	50.2	7.16
(2)	DSM: Viterbi	37.7	51.4	1.00
(3)	DSM: DC-Vite	**40.2**	**54.5**	2.67
(4)	(3)$\rightarrow$ PLR	47.2	**62.4**	3.41
(5)	(3) + PLR ($\alpha = 0.8$)	**49.3**	62.4	3.41

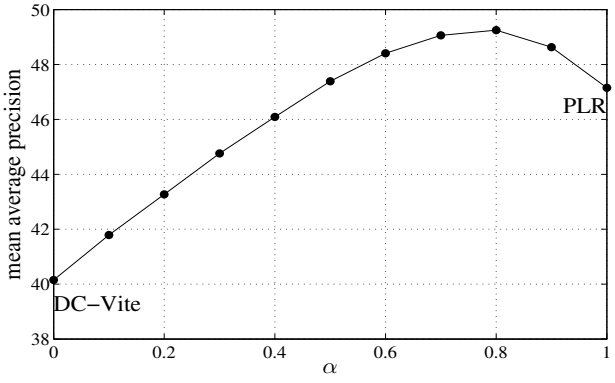

Fig. 2. Mean average precision with respect to different interpolation weight α.

5. EXPERIMENTS

5.1. DSM by DC-Vite

We evaluated the detection performance of the proposed document state matching (DSM) approach, using both Viterbi and the proposed DC-Vite as presented in Section 2. The MAP results are respectively listed in rows (2)–(3) in Table 2. The main difference between DC-Vite and Viterbi is the duration constraint. Applying the duration constraint in Viterbi decoding improved the MAP from 37.7% to 40.2%, showing that the speaking rate ratio constraint was an important issue in the matching process. Compared with the baseline segmental DTW approach (row(1)), there was a 2.7% absolute MAP improvement for DC-Vite with 2.7$\times$ times efficiency gain. As mentioned previously in Section 2.2, the number of local distance calculations for DTW is equal to the number of query frames times the number of frames in all documents. For DC-Vite, the number of likelihood $\Pr(\mathbf{x}_i|q_j)$ calculations is equal to the product of the number of query frames and the number of Gaussians in the subword-like pattern models. There were 1.49×10^6 frames in the document archive but only 2713 Gaussians in the models as shown in Table 1. The number of local distance/likelihood computations differ by more than a factor of 700. This explains the reduction in CPU time. Although the efficiency gain was not large on our experimental data set (only 4.1 hours), the discovered models showed the ability to represent speech archive without much loss of acoustic information. We expect the search time to be essentially independent of number of documents by indexing the archive with such word-like patterns in the future.

5.2. PLR

We report the results with the pseudo likelihood ratio (PLR) approach in the second stage listed in rows (4)–(5) of Table 2. From the first stages, a ranking list of hypothesized regions was returned by DC-Vite. We assume only the top K regions were competitive and to be reranked by the PLR approach, where $K = 1000$. We took as pseudo relevant examples the top $|\mathcal{R}|$ hypothesized regions obtained in the first stage, ranked with decreasing likelihood generated by DC-Vite, where $|\mathcal{R}| = 7$ in our experiments. Excluded from the pseudo relevant examples, $|\mathcal{I}|$ hypothesized regions were randomly selected from the top K competitive hypothesized regions to form set $\mathcal{I}$ and to train $\Lambda_{\mathcal{I}}$, where $|\mathcal{I}| = 50$ in our experiments. The decay parameter λ for training $\Lambda_{\mathcal{R}}$ was set to 0.25. After the two models were trained for each query, the Viterbi algorithm was performed on every hypothesized region in the top K list using both models. For each hypothesized region $\mathbf{H}$, the relevant score was set to be the log likelihood ratio of $\mathbf{H}$ to $\Lambda_{\mathcal{R}}$ and $\Lambda_{\mathcal{I}}$. We reranked the top K hypothesized regions according to their log likelihood ratios, and froze all other lower ranked hypothesized regions in the list. The results are listed in row (4). We see that performing PLR in the second stage has a 7.0% absolute MAP improvement and a 7.9% absolute P@10 improvement over the DC-Vite approach. It shows that with enough pseudo relevant examples, we can more precisely characterize the distribution and trajectory of the query using a whole word HMM. Also, emphasizing the likelihood ratio of discriminative signal parts that separate the true spoken terms from competitive false accepted regions is very effective for accurate detection.

We also considered relevance score integration for DC-Vite and PLR. After normalizing the DC-Vite scores and PLR scores to zero mean and unit variance, the fused score for each hypothesized region in the top K list was computed as a linear interpolation of the scores obtained from the two approaches with an interpolation weight α. The top K hypothesized regions were reranked according to the interpolated weight. The results are equivalent to DC-Vite in row (3) of Table 2 if $\alpha = 0$, and $\alpha = 1$ produces the result of PLR in row (4). Fig. 2 shows the MAP with different alpha, with optimal $\alpha = 0.8$. It is clear that the combination can achieve better detection performance than each approach alone.

6. CONCLUSION

We propose a two-stage model-based approach for unsupervised QbE STD using a set of models generated by automatic acoustic pattern discovery. By representing spoken documents as the states in the acoustic patterns, we can match query frames to document states without calculating frame-wise distance as in DTW. The proposed duration-constrained Viterbi approach can eliminate unrealistic speaking rate distortion thus outperforms conventional Viterbi decoding. The proposed pseudo likelihood ratio approach further improves the detection performance by modeling the distribution of pseudo relevant examples and competitive pseudo irrelevant examples. Experimental results show a 11.8% absolute MAP improvement over the segmental DTW approach, which suggests the feasibility of indexing speech data with automatically discovered acoustic patterns. The proposed approach highlights the potential of leveraging well-developed HMM-based speech processing techniques for model-based approaches for zero-resource STD.

7. REFERENCES

[1] NIST, "The spoken term detection (STD) 2006 evaluation plan, 10th ed.," http://www.nist.gov/speech/tests/std.

[2] David R. H. Miller, Michael Kleber, Chia-Lin Kao, Owen Kimball, Thomas Colthurst, Stephen A. Lowe, Richard M. Schwartz, and Herbert Gish, "Rapid and accurate spoken term detection," in *Proc. INTERSPEECH*, 2007.

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

[4] R. Wallace, R. Vogt, and S. Sridharan, "A phonetic search approach to the 2006 NIST spoken term detection evaluation," in *Proc. INTERSPEECH*, 2007.

[5] Makoto Terao, Takafumi Koshinaka, Shinichi Ando, Ryosuke Isotani, and Akitoshi Okumura, "Open-vocabulary spoken-document retrieval based on query expansion using related web documents," in *Proc. INTERSPEECH*, 2008.

[6] Yi-Cheng Pan and Lin-shan Lee, "Performance analysis for lattice-based speech indexing approaches using word and subword units," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, August 2010.

[7] Murat Saraclar and Richard Sproat, "Lattice-based search for spoken utterance retrieval," in *Proc. HLT-NAACL*, 2004.

[8] Carolina Parada, Abhinav Sethy, and Bhuvana Ramabhadran, "Query-by-example spoken term detection for OOV terms," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2009.

[9] Timothy J. Hazen, Wade Shen, and Christopher White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2009.

[10] Lou Boves, Rolf Carlson, Erhard W. Hinrichs, David House, Steven Krauwer, Lothar Lemnitzer, Martti Vainio, and Peter Wittenburg, "Resources for speech research: present and future infrastructure needs," in *Proc. INTERSPEECH*, 2009.

[11] Arun Kumar, Nitendra Rajput, Dipanjan Chakraborty, Sheetal K. Agarwal, and Amit A. Nanavati, "WWTW: the world wide telecom web," in *Proc. of the 2007 workshop on Networked systems for developing regions*, 2007.

[12] Florian Metze, Nitendra Rajput, Xavier Anguera, Marelie Davel, Gravier Guillaume, Charl van Heerden, Gautam V. Mantena, Armando Muscariello, Kishore Prahallad, Igor Szöke, and Javier Tejedor, "The spoken web search task at MediaEval 2011," in *ICASSP*, 2012.

[13] Chun-an Chan and Lin-shan Lee, "Unsupervised hidden Markov modeling of spoken queries for spoken term detection without speech recognition," in *Proc. INTERSPEECH*, 2011.

[14] Michael A. Carlin, Samuel Thomas, Aren Jansen, and Hynek Hermansky, "Rapid evaluation of speech representations for spoken term discovery," in *Proc. INTERSPEECH*, 2011.

[15] Yaodong Zhang and James R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2009.

[16] Marijn Huijbregts, Mitchell McLaren, and David van Leeuwen, "Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection," in *ICASSP*, 2011.

[17] Haipeng Wang, Cheung-Chi Leung, Tan Lee, Bin Ma, and Haizhou Li, "An acoustic segment modeling approach to query-by-example spoken term detection," in *ICASSP*, 2012.

[18] Chun-an Chan and Lin-shan Lee, "Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping," in *Proc. INTERSPEECH*, 2010.

[19] Chun-an Chan and Lin-shan Lee, "Integrating frame-based and segment-based dynamic time warping for unsupervised spoken term detection with spoken queries," in *ICASSP*, Prague, May 2011.

[20] Yaodong Zhang and James R. Glass, "A piecewise aggregate approximation lower-bound estimate for posteriorgram-based dynamic time warping," in *Proc. INTERSPEECH*, 2011.

[21] Yaodong Zhang and James R. Glass, "Fast spoken query detection using lower-bound dynamic time warping on graphical processing units," in *ICASSP*, 2012.

[22] Aren Jansen and Benjamin Van Durme, "Indexing raw acoustic feature for scalable zero resource search," in *Proc. INTERSPEECH*, 2012.

[23] Cheng-Tao Chung, Chun-an Chan, and Lin-shan Lee, "Unsupervised discovery of linguistic structure including two-level acoustic patterns using three cascaded stages of iterative optimization," in *ICASSP*, 2013.

[24] M. Ostendorf, V. Digalakis, and O. A. Kimball, "From HMMs to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 360–378, 1995.

[25] Myoung-Wan Koo, Chin-Hui Lee, and Biing-Hwang Juang, "Speech recognition and utterance verification based on a generalized confidence score," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 821–832, 2001.

[26] Ellen M. Voorhees, "Overview of the TREC 2006," in *TREC*, 2006.