A Fast Query-by-Example Spoken Term Detection for Zero Resource Languages

Karthik Pandia D S
Dept. of Computer Science and Engg.
Indian Institute of Technology Madras
Email: pandia@cse.iitm.ac.in

Saranya M S
Dept. of Computer Science and Engg.
Indian Institute of Technology Madras
Email: saranms@cse.iitm.ac.in

Hema A Murthy
Dept. of Computer Science and Engg.
Indian Institute of Technology Madras
Email: hema@cse.iitm.ac.in

***Abstract*—This paper presents a novel two-pass dynamic time warping (DTW) approach to build Query-by-Example Spoken Term Detection (QbE-STD) system for Zero Resource Languages. An unconstrained-endpoint dynamic time warping (UE-DTW) algorithm is used to locate the query term occurrences in a long conversational audio. The proposed approach uses a segmental DTW, wherein search is carried out only at syllable boundaries. This reduces the search complexity by 9 times compared to conventional sliding window DTW. The first pass of the proposed method uses a minimum set of templates for a keyword to search through the segmented audio. New templates are identified after the first pass. In the second pass, the initial templates along with the new templates identified in the first pass are used to search for the keyword occurrences. A novel score normalization technique is also proposed, in which the syllables constituting the keyword are used for normalization. The performance of the proposed two-pass system is shown to be better than the single pass systems. The proposed score normalization technique further improves the overall detection results.**

I. Introduction

The proliferation of video websites has led to a significant increase in the amount of audio and video content on the web. Nevertheless, audio/video is mined using the associated text tags. Querying audio/video using spoken keywords is an important task, especially given the growing availability of video lectures (MIT-OCW [1], Stanford [2], CMU [3], NPTEL [4], etc). This work attempts to solve the spoken term detection (STD) problem using the audio templates on lecture videos. Some of the best present day STD methods use Automatic Speech Recognition (ASR) systems, which demand a huge amount of annotated speech data to train appropriate models [5]–[7]. Even if annotations are available, use of some of the words in the vernacular leads to the out-of-vocabulary (OOV) problem. Moreover, a high word-level accuracy of modern ASR systems relies on effective language models [8], which may not be always available. To address these issues, a phone-based recognizer was used in [9] and [10]. But poor phone level accuracy can lead to significant errors.

STD is a more challenging task when there is little or no resource available. In such a scenario, the QbE is the most suitable technique. QbE-STD is the task of searching audio content within a considerably longer audio file by using an audio query. This has been an important task owing to the increase in the amount of audio content in the web. The QUESST, a QbE-STD evaluation campaign being conducted annually by MediaEval [11], focuses on this problem. In the subsequent sections, an example query is termed as a *query template* and the audio file in which the query is searched is referred to as *search file*. The proposed work aims to build a QbE-STD system that works in a zero resource environment natural speech data. For this purpose, the NPTEL lecture corpus is chosen. The videos are highly accented technical speech with a lot of out-of-vocabulary (OOV) words. The search file is conversational speech, which makes the task challenging.

The three main parameters to evaluate an STD system are Response time, performance and flexibility [12]. DTW based keyword spotting using Mel Frequency Cepstral Coefficients (MFCC) was first proposed by Sakoe and Chiba [13]. Huge response time due to the exhaustive frame-wise search is a major drawback of the conventional DTW based STD system, especially when the search file is of long duration. Our approach uses a Segmental DTW to reduce the time complexity of the search process. Segmental DTW proposed in [14] finds multiple alignment paths corresponding to the diagonal region by applying the band constraints to find matching acoustic patterns between the query and the test utterance. The authors have also extended their work to Gaussian posterior based features [15]. Even though the method is quite useful for the unsupervised pattern discovery, it does not help to reduce the search space. Another variant of SDTW called Segmental Locally Normalized DTW (SLNDTW) was proposed in [16], where the starting point of the search segments are obtained by comparing each cell of the distance matrix with its left neighbor. Despite the fact that the method helps to reduce the search space, it computes the distance matrix for the entire test segment. In [17], a segmental DTW was used where the segments are obtained by grouping the acoustically similar feature vectors into one representative vector. The authors have refined the segmental DTW using hierarchical agglomerative clustering [18]. Memory Efficient Subsequence-DTW (MES-DTW) proposed in [19] identifies the ending point from the entire DTW matrix and traces back to identify the starting point. Most of the SDTW variants perform some kind of preprocessing to segment the search file.

This paper uses a segmental DTW (SDTW) proposed in [20] to reduce the search space and thereby improving the

978-1-5090-1746-1/16/$31.00 © 2016 IEEE

response time of the system. Here the boundaries corresponding to the syllables are identified using a Group Delay based segmentation (GDS) algorithm [21], [22]. The search for a query template is performed on every syllable beginning as shown in Figure 1. The segments at every syllable boundary are termed as *test segments*. The query template is matched with the search file using a sliding window approach on the test segments. One syllable boundary (test segment) is moved at a time. Apart from this, a two-pass system is proposed. Here the system aims at finding new templates automatically from the search segment with the help of a minimal number of initial set of templates. The paper also proposes a score normalization technique to improve the word spotting performance of the system. Rest of the paper is organized as follows: Section II gives a brief description of the database. The proposed method is detailed in Section III. Section IV explains the experimental setup, description of the different systems used, and an analysis of the results. Finally, Section V concludes the paper.

II. Database description

The National Programme on Technology Enhanced Learning (NPTEL), is a project funded by the Ministry of Human Resource Development (MHRD), India. Under this project Indian Institutes of Technology (IIT Bombay, Delhi, Guwahati, Kanpur, Kharagpur, Madras and Roorkee) and Indian Institute of Science, Bangalore have worked together to develop the web and video based material for basic undergraduate science and engineering courses. The video materials are generated by recording the lectures in an air-conditioned classroom environment with an omni-directional microphone. There are approximately 110 courses offered by different professors across the country. A course contains 40 lectures and each lecture is primarily delivered by a single speaker addressing the audience. Duration of a lecture is approximately one hour. The course lectures tend to have relatively small vocabularies where subject specific words are used frequently. These lectures are highly vernacular in nature and do not have precise transcriptions. Though these lectures are in English, they are highly accented by the speaker's native tongue. From the NPTEL corpus, 10 courses are used for the experimental evaluation. 60 keywords from ten different courses are chosen for the experiment.

III. Proposed system

As mentioned earlier, a GDS based segmental DTW proposed in [20] is used to improve the search time. In the GDS algorithm, by tuning a parameter called window scale factor (WSF), the segmentation boundaries can be varied from phonemes to words. In the proposed algorithm, the WSF is tuned to overestimate the number of syllable boundaries, nevertheless, it effectively reduces the search space. Syllable boundaries are preferred based on the fact that a word always begins with a syllable. Word level segmentation may not be appropriate, as occasionally a boundary could be missed.

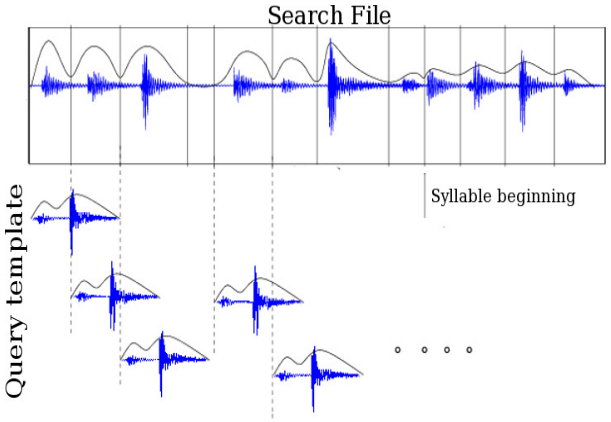


Fig. 1: Illustration of segmental search of query template on the syllable boundaries of the search file using GDS

A. Two-pass QbE-STD system

The proposed system uses a two-pass search strategy. In the first pass, query matches from the search file are obtained, and fed back as new query instances for the second pass. To illustrate the effectiveness of using a two-pass search and to ensure the robustness of the system, the total number of initial query templates are restricted to three. These three initial templates are manually chosen and extracted from the search file. During the first pass, the query occurrences are searched at the syllable boundaries of the search file using the initial set of templates.

As the search file is much longer than the query, an unconstrained version of DTW is used. One end of the test segment is fixed and the other end is open, making it an unconstrained-endpoint DTW (UE-DTW) Figure 1. Three new templates are obtained from the search performed on the first path. These three templates correspond to the segments with top three DTW scores arranged in increasing order. These three new segments along with the initial query templates make a total of six query templates. In the second pass, SDTW is applied on the search file using all the six query templates. One DTW score for each test segment is computed using six templates. The geometric mean of the six DTW scores is used as the final score (Equation 1). The geometric mean is found to work better than the arithmetic mean. By using geometric mean, a good match of one template is sufficient for the overall score to be very small.

$$D(K, T_s) = \left[\prod_{i=1}^{N} \text{UE-DTW}(Q_i, T_s)\right]^{\frac{1}{N}} \qquad (1)$$

where $D(K, T_s)$ is the final score between the keyword K and the test segment T_s. Q_i is the i^{th} query template, and N is the total number query templates.

B. Syllable based score normalization

Score normalization is a crucial task for any verification system. It helps to bring the scores to a common base and to reduce the overlap between the targets and non-target scores.

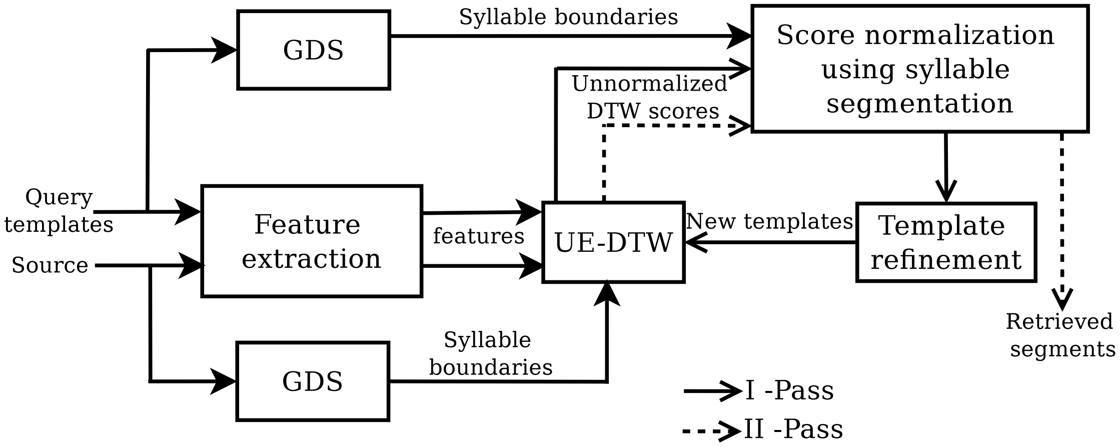


Fig. 2: Proposed two-pass Query-by-Example Spoken Term Detection system (2P-QbE-STD)

In the case of the QbE-STD task, the query templates are of different lengths. This has to be handled using a score normalization technique. Several normalization techniques are available for DTW based systems [23]–[26]. One of the classic normalization techniques is the length normalization technique proposed in [27]. Here the normalization function is defined according to the step pattern used in the DTW algorithm. But this length normalization proposed for the isolated word recognition does not seem to work for continuous speech. Usage of unnormalized scores shows a better performance than the length normalization. In [20], a score normalizing technique is used in which the scores of the test segments are normalized with respect to an arbitrary set of anti-keywords.

In this paper, we propose a normalization technique that uses the syllables constituting the query template. For every pair of test segment and the query template, DTW score is computed and normalized with respect to the combined DTW scores between the test segment and the syllables constituting the query as given in Equation 2.

$$D(Q,T_s) = \frac{\text{UE-DTW}(Q,T_s)}{\frac{1}{N}\sum_{i=1}^{N}\text{DTW}(T_s,S_i)} \quad (2)$$

where $D(Q,T_s)$ is the syllable normalized distance between the query template Q and the test segment T_s. N is the total number of syllables in the query and S_i refers to the i^{th} syllable from the syllable sequence constituting the query. This normalization improves the QbE-STD performance by reducing the number of false alarm and missed detection. The performance of different QbE-STD systems with and without normalization are compared in Section IV-B.

IV. Experiments and result analysis

The proposed method is implemented and tested on a subset of NPTEL lecture corpus as discussed in Section II. Various QbE-STD systems used for the experiments along with the proposed systems are listed below.

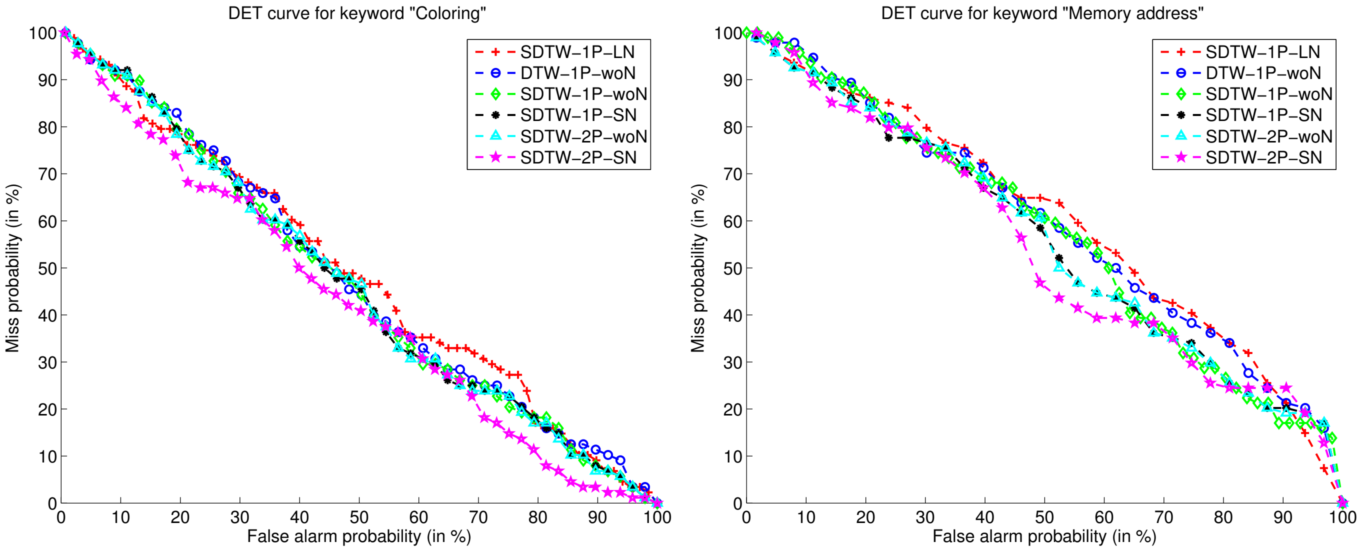
- DTW-1P-woN: conventional sliding window DTW computation without normalization
- SDTW-1P-woN: one-pass segmental DTW without normalization
- SDTW-1P-LN: one-pass segmental DTW with length normalization
- SDTW-1P-SN: one-pass segmental DTW with syllable normalization
- SDTW-2P-woN: two-pass SDTW systems without normalization
- SDTW-2P-SN: two-pass SDTW systems with syllable normalization

A. Experimental setup

The flow of the proposed two-pass system is illustrated in Figure 2. Three templates are chosen for every query keyword. MFCC features are extracted for all the query templates and the search file. GDS is applied on both, templates and search file, to get the syllable boundaries. In the first pass, unnormalized DTW scores are computed between the query and the test segments (Equation 1). These scores are normalized using syllable normalization (Equation 2). Three new templates are chosen from the retrieved segments. In the second pass, these three templates along with the initial templates are used to get new DTW scores. These scores are again syllable normalized to obtain final scores.

B. Result analysis

The performances of all the systems used for the experiment, in terms of Equal Error Rate (EER) are shown in Table I. The Detection Error Trade-off (DET) curves of two keywords "coloring" and "memory address" for various systems used in the experiments are shown in Figure 3. The computational times of QbE-STD systems are compared in terms of the average time taken to search a keyword. The length normalization procedure in [27] with single pass performs worser than the system that does not use any score normalization. Using syllable normalization on single pass system yields better performance than the unnormalized system. Despite the fact that the EER of the sliding window DTW system is slightly (0.17%) better than the single pass syllable



(a)

(b)

Fig. 3: Sub-figures (a) and (b) shows the DET curves of keywords “coloring” and “memory address” for various QbE-STD systems

normalization system, the computational time of the sliding window DTW system is 4 times more. The proposed two-pass systems (SDTW-2P-woN and SDTW-2P-SN) are shown to be superior to all single pass systems. Even though the syllable normalization gives marginal improvement in EER when compared to single pass systems, the improvement is significant in case of the two-pass SDTW systems. The high EER values obtained is due to a large number of test segments used for the evaluation. The total number of test cases for the evaluation is around 5 million for 60 keywords. Moreover, the partially matched words are also considered as false alarms, leading to the increase in overlap between the true and impostor scores.

TABLE I: Average EERs of various QbE-STD systems used for the evaluation

System	**Average EER (in%)**
SDTW-1P-LN	51.9
DTW-1P-woN	51.4
SDTW-1P-woN	51.8
SDTW-1P-SN	51.5
SDTW-2P-woN	51.1
SDTW-2P-SN	**50.2**

TABLE II: Average response time (T_A) for a template search on a 60 min long test file

System	T_A **(in sec)**
DTW-1P-woN	455.2
SDTW-1P-woN	50.2
SDTW-1P-SN	101.7

The segmental DTW using the GDS helps to reduce the search time drastically. For search file (course lectures) of one hour duration, with a sampling rate of 16 kHz and a frame shift of 160 samples, the number of feature vectors to be processed by sliding window DTW is about 0.35 million. On the other hand, the total number of syllable segments in a lecture is around $30,000$. This reduces the search space drastically. Window scale factor is an important parameter for the GDS. Large WSF will lead to an overestimate of the boundaries and small value will lead to the missing of syllable boundaries. The average time taken to complete a query search for the one-pass systems are shown in Table II. It can be seen that the sliding window DTW system took 455.2 seconds to search through a search file whereas the SDTW-1P-woN took 50.2 seconds. Thus, computational time is reduced by a factor of 9. The syllable normalization based SDTW (SDTW-1P-SN) took an average search time of 101.7 seconds, which is approximately twice the time taken by the SDTW-1P-woN. This is because of the additional DTW computations performed for each syllable segment. Similar results are observed for the two-pass systems. A serious drawback of the proposed system is that the initial templates have to be manually chosen. There is also a considerable number of miss in the syllable boundaries is seen in case of fast utterances. Apart from using group-delay, other acoustic cues can be used to avoid missing boundaries.

V. Conclusion

We have proposed a two-pass QbE-STD, which is used in tandem with the existing group-delay based segmentation algorithm. The proposed system with a two-pass DTW along with the syllable normalization shows a significant improvement in performance compared to the simple single pass systems. As there is no training process involved and additional query segments are extracted from the test file, the system can be used in an unsupervised scenario, especially for zero resource languages.

VI. Acknowledgement

The authors would like to acknowledge the Defence Research and Development Organization, India for funding the research under the project CSE1314142DRDOHEMA.

References

[1] MIT courses, "http://ocw.mit.edu/," .

[2] Stanford Courses, "http://online.stanford.edu/courses," .

[3] CMU courses, "http://oli.cmu.edu/," .

[4] NPTEL courses, "http://onlinecourses.nptel.ac.in," .

[5] Yi Wu, Rong Zhang, and A. Rudnicky, "Data selection for speech recognition," in *IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, Dec 2007, pp. 562–565.

[6] A. Nagroski, L. Boves, and H. Steeneken, "In search of optimal data selection for training of automatic speech recognition systems," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Nov 2003, pp. 67–72.

[7] Arkadiusz Nagórski, Lou Boves, and Herman JM Steeneken, "Optimal selection of speech data for automatic speech recognition systems.," in *3rd Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2002.

[8] Scott Novotney, Richard Schwartz, and Jeff Ma, "Unsupervised acoustic and language model training with small amounts of labelled data," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2009, pp. 4297–4300.

[9] D.A. James and S.J. Young, "A fast lattice-based approach to vocabulary independent wordspotting," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr 1994, vol. 1, pp. 377–380.

[10] K. Thambiratnam and S. Sridharan, "Dynamic match phone-lattice searches for very fast and accurate unrestricted vocabulary keyword spotting," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2005, vol. 1, pp. 465–468.

[11] Media Eval, "http://www.multimediaeval.org," .

[12] Moyal Ami, Aharonson Vered, Tetariy Ella, and Gishri Michal, *Phonetic Search Methods for Large Speech Databases*, Springer, New York, 2013.

[13] Hiroaki Sakoe and Seibi Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 1, pp. 43–49, 1978.

[14] A.S. Park and J.R. Glass, "Unsupervised pattern discovery in speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 186–197, Jan 2008.

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

[16] O Muscariello, Guillaume Gravier, Frdric Bimbot, Inria Rennes, and Bretagne Atlantique, "Audio keyword extraction by unsupervised word discovery," in *10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2009.

[17] Chun-an Chan and Lin-Shan Lee, "Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping," in *11th Annual Conference of the International Speech Communication Association INTERSPEECH*, Makuhari, Chiba, Japan, September 26-30 2010, pp. 693–696.

[18] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 5652–5655.

[19] X. Anguera and M. Ferrarons, "Memory efficient subsequence dtw for query-by-example spoken term detection," in *IEEE International Conference on Multimedia and Expo (ICME)*, July 2013, pp. 1–6.

[20] Srikanth R Madikeri and Hema A Murthy, "Acoustic segmentation using group delay functions and its relevance to spoken keyword spotting," in *Text, Speech and Dialogue*. Springer, 2012, pp. 496–504.

[21] T Nagarajan, Hema A Murthy, and Rajesh M.Hegde, "Group delay based segmentation of spontaneous speech into syllable-like units," in *ISCA and IEEE Worksop on Spontaneous Speech Processing and Recognition*, April 2003.

[22] T Nagarajan, V Kamakshi Prasad, and Hema A Murthy, "The minimum phase signal derived from the magnitude spectrum and its application to speech segmentation," in *Sixth Biennial Conference of Signal Processing and Communications*, July 2001.

[23] A. Fischer, M. Diaz, R. Plamondon, and M.A. Ferrer, "Robust score normalization for dtw-based on-line signature verification," in *13th International Conference on Document Analysis and Recognition (ICDAR)*, Aug 2015, pp. 241–245.

[24] Haipeng Wang and Tan Lee, "Cuhk system for the spoken web search task at mediaeval 2012.," in *MediaEval*. 2012, vol. 927, CEUR-WS.org.

[25] Igor Szke, Luks Burget, Frantisek Grzl, and Lucas Ondel, "But sws 2013 - massive parallel approach.," in *MediaEval*. 2013, vol. 1043 of *CEUR Workshop Proceedings*, CEUR-WS.org.

[26] I. Szoke, L. Burget, F. Grezl, J.H. Cernocky, and L. Ondel, "Calibration and fusion of query-by-example systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 7849–7853.

[27] C. Myers, L. Rabiner, and A.E. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 6, pp. 623–635, Dec 1980.