



Unsupervised Spoken-Term Detection with Spoken Queries Using Segment-based Dynamic Time Warping

Chun-an Chan, Lin-shan Lee

National Taiwan University, Taipei, Taiwan, R.O.C.

chunananchan@gmail.com, lslee@gate.sinica.edu.tw

Abstract

Spoken term detection is important for retrieval of multimedia and spoken content over the Internet. Because it is difficult to have acoustic/language models well matched to the huge quantities of spoken documents produced under various conditions, unsupervised approaches using frame-based dynamic time warping (DTW) has been proposed to compare the spoken query with spoken documents frame by frame. In this paper, we propose a new approach of unsupervised spoken term detection using segment-based DTW. Speech signals are segmented into sequences of acoustically similar segments using hierarchical agglomerative clustering, and a DTW procedure is formulated for segment sequences along with the clustering tree structures. In this way, the number of highly redundant parameters can be reduced, and the relatively unstable feature vectors can be replaced by more stable segments which describe the sequence of vocal track stages during the uttering process. Preliminary experiments indicate a high reduction of computation time as compared to frame-based DTW, although the slightly degraded detection performance implies much room for further improvements.

Index Terms: spoken term detection, segment-based speech processing, dynamic time warping

1. Introduction

As the multimedia content on the Internet increases exponentially today, spoken term detection becomes very important for retrieval of spoken or multimedia documents. The goal of spoken term detection is to find in the document archive the desired spoken term for a query entered [1]. In this paper, we consider the scenario that the query entered is also a spoken term. Most of the techniques for spoken term detection require a speech recognizer to transform the speech signals to language units, such as words, syllables, or phonemes [2, 3, 4, 5]. Matching process is then performed on the transformed unit strings or lattices and a relevance score between the query and the spoken utterance is obtained. The performance of such a retrieval system relies heavily on the performance of the speech recognizer [6]. However, it is difficult to have a set of acoustic/language models well matched to the huge quantities of spoken documents produced by many different speakers under many different acoustic conditions. Some researchers thus proposed the approach of unsupervised spoken-term detection, in which the signals are transformed to sequences of feature vectors and the similarity between the spoken term and the spoken documents is computed via template matching techniques such as dynamic time warping (DTW) [7, 8].

Gaussian posteriorgrams were used as feature vectors for segmental DTW earlier [7], and in other works a speech rec-

ognizer was also used to generate the phonetic posteriorgrams used as feature vectors in modified DTW [8]. These works focused on the frame-based feature vectors describing the fine structures of the speech signals. However, the number of frames for a spoken query or test utterance can be as many as several thousands. The computation requirements will be very high practically. Also the frame-based feature does not carry speech information of longer term. In this work, we present the initial results for a slightly different approach, segment-based DTW. We first perform segmentation on the sequence of feature vectors to create a sequence of segments (groups of consecutive acoustically similar frames) for further processing. In this way, the number of highly redundant parameters can be reduced, and the relatively unstable feature vectors can be replaced by more stable segments which describe sequences of vocal track stages during the uttering process. The segments are actually the fundamental units to construct the phonemes, words and so on. We formulate the segment-based DTW and compare it with the frame-based DTW. The initial experimental results show that the computation time can be highly reduced but with a slight degradation in detection performance reduction. Improved techniques with better detection performance is in good progress.

2. Segmentation of feature vector sequences

The query term and test utterances are first converted to spectrogram. We then split these utterances into connected segments, or acoustically similar groups of neighboring vectors [9].

2.1. Segmentation by hierarchical agglomerative clustering

Let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ be a sequence of feature vectors, where each \mathbf{x}_t is a vector, and $S = (t_0, \dots, t_L)$ be the segment boundaries with $0 = t_0 < t_1 < \dots < t_L = N$. So the l -th segment $(\mathbf{x}_{t_{l-1}+1}, \dots, \mathbf{x}_{t_l})$ ends at t_l and L is the total number of segments. For a feature vector sequence \mathbf{X} and a segment boundary set S , the sum of squared error is

$$f(\mathbf{X}, S) = \sum_{l=1}^L \sum_{t=t_{l-1}+1}^{t_l} \|\mathbf{x}_t - \mathbf{m}_l\|^2, \quad (1)$$

where

$$\mathbf{m}_l = \frac{1}{t_l - t_{l-1}} \sum_{t=t_{l-1}+1}^{t_l} \mathbf{x}_t \quad (2)$$

is the mean vector of the l -th segment. When the total number of segments L is given, hierarchical agglomerative clustering (HAC) can be used to find S that minimize the function $f(\mathbf{X}, S)$ [9]. The initial segment boundary set S_0 is $\{0, 1, 2, \dots, N\}$. At each iteration i , two consecutive segments that minimize the

10.21437/Interspeech.2010-262

Table 1: Evaluation of unsupervised segmentation on TIMIT.

β	Tolerance	Precision	Recall	F-measure
0.15	10ms	0.57	0.69	0.63
	20ms	0.68	0.82	0.74
	30ms	0.71	0.86	0.78
0.50	10ms	0.68	0.58	0.62
	20ms	0.82	0.70	0.75
	30ms	0.86	0.73	0.79

loss function $\mathcal{L}(i)$ will be merged to a new segment and the boundary set S_{i-1} is changed to S_i by deleting a boundary t_j in S_{i-1} , where

$$\mathcal{L}(i) = f(\mathbf{X}, S_i) - f(\mathbf{X}, S_{i-1}). \quad (3)$$

The algorithm stops when there are L segments left. This greedy algorithm does not guarantee optimal solution, but with a time complexity of $O(L \log N)$, which is good for faster implementation.

2.2. Number of segments L

In this work, the number of segment L is unknown and need to be decided. We perform HAC on the feature vectors until there is only one segment left, which is the whole utterance. If there are L acoustically similar segments in the utterance and these L segments appear in the HAC process, merging subsegments within these segments should give relatively small $\mathcal{L}(i)$ as in equation 3 and merging two of these L segments should give large $\mathcal{L}(i)$. Hence $\mathcal{L}(i)$ should boost up when there are L segments left, or the first $N - 1 - L$ merge losses should be relatively small compared with the last L merge losses. In general, N is about ten times larger than L , so the mean of $\mathcal{L}(i)$, $mean(\mathcal{L}(i))$, is only slightly higher than the mean of the first $N - 1 - L$ losses. So we define a threshold to stop the merging process by taking the dynamic range of $\mathcal{L}(i)$ into consideration. We let the threshold $\lambda = mean(\mathcal{L}(i)) + \beta \cdot std(\mathcal{L}(i))$, where $std(\mathcal{L}(i))$ is the standard deviation of $\mathcal{L}(i)$ and β is a parameter. The segments will be finer if β is smaller. The L segments obtained in this way are referred to as ‘‘basic segments’’ of the utterance in the following.

We performed an experiment on 6300 TIMIT utterances to see if this segmentation can detect phonetic boundaries. The precision, recall and F-measure of segmentation boundaries compared with phone boundaries are shown in Table 1 with $\beta = 0.15$ and 0.5. The F-measure of these two values of β are similar, while a smaller β gives higher recall rate, but also more false alarms. We choose $\beta = 0.15$ because a false rejection causes much more problems than a false acceptance in the following processing.

3. Dynamic time warping (DTW)

In this section, we first formulate the previously developed frame-based DTW with modified constraints and objectives, based on which we then formulate the segment-based DTW proposed here.

3.1. Frame-based DTW

The frame-based DTW is to match two feature vector sequences of different lengths for distance computation. Let $\mathbf{Q} = (q_1, \dots, q_{N_Q})$ be the spoken query feature sequence and $\mathbf{Y} = (y_1, \dots, y_{N_Y})$ a test utterance feature sequence, the goal

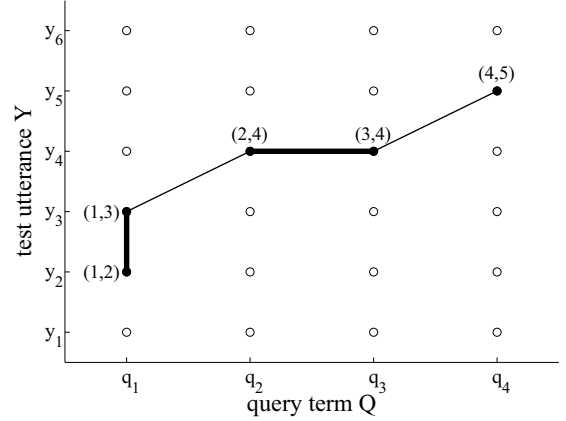


Figure 1: A warping path that matches (q_1, \dots, q_4) with (y_2, \dots, y_5) . The bold lines are maximal horizontal and vertical subpaths.

is to find a warping path, or a matched index pair sequence,

$$P = \{(i_1, j_1), (i_2, j_2), \dots, (i_K, j_K)\}, \quad (4)$$

where i_k and j_k are frame indices for Q and for Y respectively, with

$$1 = i_1 \leq i_2 \leq \dots \leq i_K = N_Q, \quad (5)$$

$$1 \leq j_1 \leq j_2 \leq \dots \leq j_K \leq N_Y, \quad (6)$$

$$i_{k+1} - i_k \leq 1, j_{k+1} - j_k \leq 1, \quad (7)$$

such that a predefined distance $D(Q, Y, P)$ is minimized. The constraint (5) forces the whole spoken query to be matched, while constraint (6) places no such limitation on the matched test utterance. Figure 1 shows an example warping path that matches $Q = (q_1, \dots, q_4)$ with (y_2, \dots, y_5) .

There are different ways to evaluate the distance measure $D(Q, Y, P)$ and different constraints on the warping path P . In segmental DTW [7, 10],

$$D(Q, Y, P) = \sum_{k=1}^K d(q_{i_k}, y_{j_k}), \quad (8)$$

where $d(\cdot)$ is the distance measure between two Gaussian posteriorgram vectors, and the constraints on P is

$$|(i_k - i_1) - (j_k - j_1)| < R, \quad (9)$$

where R is a parameter preventing an overly large temporal skew between Q and Y . The path that gives minimum distance can be found by dynamic programming in $O(N_Q \cdot N_Y)$ time.

The above suffers from speaking rate distortion in spoken term detection. With equation (8), a test sequence with slower speaking rate will have larger distance. For a fair comparison between different test utterances, the distance of those frames in Y matched to the same frame in Q should be averaged [8]. Moreover, constraint (9) allows only R -frame differences along the path. So a warping path that matching Q with Y is not allowed when they represent the same term but the duration of Y is twice as that of Q and the length of Q is larger than R . A better constraint will be limiting the maximum number of frames that can be matched to the same feature frame. Below, we formulate the problem by adopting these two ideas. In the

example warping path in Figure 1, we see it includes a maximal vertical subpath $\{(1, 2), (1, 3)\}$, a maximal horizontal subpath $\{(2, 4), (3, 4)\}$, and a maximal subpath $\{(4, 5)\}$, which can be regarded as vertical or horizontal. This is the way to decompose a warping path P into maximal vertical or horizontal subpaths, $P = (\tilde{P}_1, \tilde{P}_2, \dots, \tilde{P}_{\tilde{K}})$, where $\tilde{P}_k, k = 1, \dots, \tilde{K}$ are maximal vertical or horizontal subpaths and a vertical subpath $\tilde{P} = \{(i_s, j_s), \dots, (i_e, j_e)\}$ is a contiguous portion of P with $i_s = i_{s+1} = \dots = i_e$, and a vertical subpath with $j_s = j_{s+1} = \dots = j_e$. A vertical/horizontal subpath is maximal if it is not a proper subpath of another vertical/horizontal subpath. With a constraint

$$(i_{k+1} - i_{k-1})^2 + (j_{k+1} - j_{k-1})^2 > 2 \quad (10)$$

on P , vertical and horizontal subpaths can be separated. If a maximal vertical subpath intersects with a maximal horizontal subpath on the end point (i_k, j_k) , this point will form a right angle on the plane and violate constraint (10). With the above, we can set the constraint

$$\left| \tilde{P}_k \right| \leq \gamma, \quad (11)$$

where $\left| \tilde{P}_k \right|$ is the length of \tilde{P}_k and γ serves as a maximum speaking rate distortion constraint. The distance of Q and Y given P is then

$$D(Q, Y, P) = \sum_{k=1}^{\tilde{K}} D(Q, Y, \tilde{P}_k), \quad (12)$$

and for any $\tilde{P}_k = \{(i_s, j_s), \dots, (i_e, j_e)\}$,

$$D(Q, Y, \tilde{P}_k) = \begin{cases} \frac{1}{|\tilde{P}_k|} \sum_{t=s}^e \|q_{i_t} - y_{j_t}\| & \text{if } i_s = \dots = i_e, \\ \sum_{t=s}^e \|q_{i_t} - y_{j_t}\| & \text{if } j_s = \dots = j_e. \end{cases} \quad (13)$$

This formulation is similar to that used previously [8], in which a rate distortion penalty different from the constraint (11) was used. Finding the optimal path with these formulation and constraints requires time complexity $O(N_Q N_Y \gamma)$.

3.2. Segment-based DTW

We now formulate DTW for any two sequences of basic segments, $\mathbf{Q} = (q_1, \dots, q_{N_Q})$ for the spoken query with q_i be the i -th basic segment in Q , and $\mathbf{Y} = (y_1, \dots, y_{N_Y})$ for a test utterance. In the example in Figure 2, the path from $(0, 0)$ to $(2, 3)$ represents the matching of a supersegment $\{q_1, q_2\}$ in Q with a supersegment $\{y_1, y_2, y_3\}$ in Y , where a supersegment may consist of several contiguous basic segments. Such a warping path P can still be represented as in equation (4), except here i_0 and j_0 stand for the indices before the start of the first supersegments in Q and in Y respectively, and for $k > 0$, i_k and j_k are the indices for the last basic segments in the k -th supersegments in Q and in Y respectively. The following constraints are similar,

$$0 = i_0 \leq i_1 \leq \dots \leq i_K = N_Q, \quad (14)$$

$$0 \leq j_0 \leq j_1 \leq \dots \leq j_K \leq N_Y, \quad (15)$$

$$i_{k+1} - i_k \leq W_Q, j_{k+1} - j_k \leq W_Y, \quad (16)$$

where W_Q and W_Y are the maximal allowed number of basic segments for supersegments in Q and Y . So the supersegment $Q(i_{k-1} + 1, i_k)$, consisting of basic segments

$(q_{i_{k-1}+1}, q_{i_{k-1}+2}, \dots, q_{i_k})$, starting from the first frame of $q_{i_{k-1}+1}$ and ending at the last frame of q_{i_k} , is matched to $Y(j_{k-1}+1, j_k)$ or $(y_{j_{k-1}+1}, y_{j_{k-1}+2}, \dots, y_{j_k})$. We further constrain the duration ratio between two supersegments to be matched. Let $f(Q(i_{k-1} + 1, i_{k+1}))$ be the total number of frames in $Q(i_{k-1} + 1, i_{k+1})$. Then the constraint is

$$\frac{1}{\bar{\gamma}} \leq \frac{f(Q(i_{k-1} + 1, i_k))}{f(Y(j_{k-1} + 1, j_k))} \leq \bar{\gamma}. \quad (17)$$

The distance of Q and Y given P is then

$$D(Q, Y, P) = \sum_{k=1}^K d(Q(i_{k-1} + 1, i_k), Y(j_{k-1} + 1, j_k)), \quad (18)$$

where the summation is over all supersegments as defined by equation (4). In the example in Figure 2, $Q = \{q_1, \dots, q_5\}$ is matched to $\{y_1, \dots, y_7\}$ by $P = \{(0, 0), (2, 3), (3, 4), (5, 6)\}$ and P is composed of 3 pairs of supersegments to be matched, $\langle Q(1, 2), Y(1, 3) \rangle$, $\langle Q(3, 3), Y(4, 4) \rangle$ and $\langle Q(4, 5), Y(5, 6) \rangle$.

We next need to define the distance between two supersegments to be used in equation (18). We use M spectrogram vectors to represent a supersegment. The value of M is the granularity of representation. Assume we are given a supersegment composed of B basic segments. Each basic segment corresponds to a subtree in the clustering tree generated by HAC. If $M > B$, we can generate a total of M subsegments by splitting the basic segment with higher loss first. If $M < B$, we can merge the basic segments with the same method described in section 2.1 until there are only M segments. Figure 3 shows an example of generating M subsegments from 3 basic segments $\{q_1, q_2, q_3\}$ via the clustering tree. When $M = 4$, q_1 , which is the basic segment with the largest merging loss, is split to q_{11} and q_{12} , producing four subsegments $\{q'_1, q'_2, q'_3, q'_4\} \triangleq \{q_{11}, q_{12}, q_2, q_3\}$. If $M = 2$, q_2 and q_3 are merged to q_4 because the loss of merging these two basic segments is the smallest and we have $\{q'_1, q'_2\} \triangleq \{q_1, q_4\}$. By averaging the feature vectors in each of these M subsegments, we get M vector representations of the supersegment.

Given a supersegment pair $\langle Q(i_{k-1} + 1, i_k), Y(j_{k-1} + 1, j_k) \rangle$, we now have two sequences of M subsegments (q'_1, \dots, q'_M) and (y'_1, \dots, y'_M) . Let $\mathbf{v}(q'_i)$ be the mean vector of q'_i , $f(q'_i)$ the number of frames in q'_i and $p(q'_i) =$

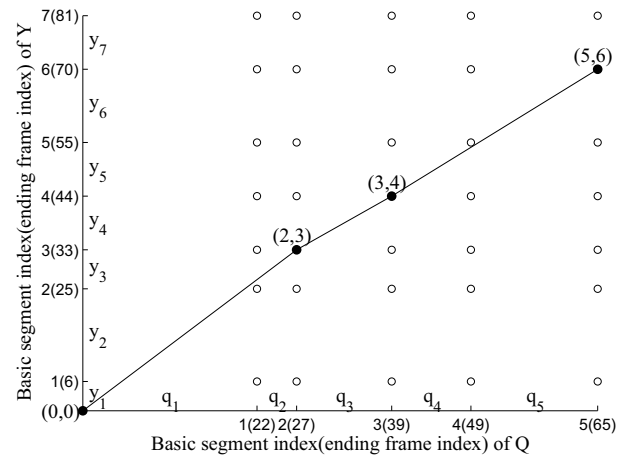


Figure 2: An example of warping path of two basic segment sequences.

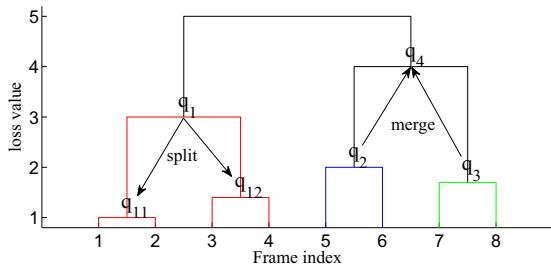


Figure 3: An example of basic segments $\{q_1, q_2, q_3\}$ split to $\{q'_1, q'_2, q'_3, q'_4\} \triangleq \{q_{11}, q_{12}, q_2, q_3\}$ and merge into $\{q'_1, q'_2\} \triangleq \{q_1, q_4\}$.

$\frac{f(q'_i)}{f(q'_1) + \dots + f(q'_M)}$, we define

$$d(Q(i_{k-1} + 1, i_k), Y(j_{k-1} + 1, j_k)) = f(Q(i_k + 1, i_{k+1})) \sum_{m=1}^M e^{\alpha |p(q'_m) - p(y'_m)|} \|\mathbf{v}(q'_m) - \mathbf{v}(y'_m)\| \quad (19)$$

where the first term $f(Q(i_k + 1, i_{k+1}))$ makes the distance proportional to the length of the query segment, and the exponential term penalizes composition differences of these two super-segments with a control variable α . This objective and constraint requires $O(N_Q N_Y W_Q W_Y M)$ times for dynamic programming.

4. Experiments

The corpus used in the experiments was Mandarin broadcast news collected in Taiwan in August and September in 2001, manually segmented into 5034 utterances. We selected 10 spoken terms as the development set and 42 as the test set, all collected from the speakers in the same corpus. All utterances containing the spoken queries were excluded from the 5034 utterances. The percentage of relevant utterances for each query ranged from 0.2% to 2.2%, averaged 0.5%, of the total test utterances. The length of the test query terms ranged from 2 to 7 syllables, with majority of 2 to 3 syllables. We used the development set to determine the parameters γ , W_Q , W_Y , $\bar{\gamma}$, and α in equations (11), (16), (17), (19). We then used these parameters to evaluate the performance of segment-based DTW and frame-based DTW with constraints and objective in equation (10)-(13). The mean average precision (MAP), precision at 5 (P@5) and precision at 10 (P@10) were used to evaluate the detection performance and the average CPU time per query for efficiency. A hit or a false alarm was evaluated on per utterance basis, so a hit was counted if the returned utterance contained the desired query.

We performed experiments on frame-based DTW as described in section 3.1 with objective and constraints in equations (10)-(13) and $\gamma = 3$, and segment-based DTW as described in section 3.2 with $W_Q = W_Y = 3$, $\bar{\gamma} = 3$, $\alpha = 0.65$, and granularity $M = 3, 4, 5, 6$. The experimental results are listed in Table 2. It is clear from Table 2 that a larger M kept more details of a segment and hence offered better detection performance but with worse efficiency. The segment-based DTW achieved the highest MAP 31.9% when $M = 5$, which was 4.1% lower than that of frame-based DTW, but the CPU time was reduced by 65.4%.

We also see that increasing the granularity M improved the detection performance, but the improvement saturated at

Table 2: Experimental results for frame-based and segment-based DTW with different granularity M .

Method	MAP	P@5	P@10	CPU time	time reduction	
frame-based	36.0	67.6	56.7	136.3s	–	
seg.-based	$M = 3$	31.2	62.9	51.2	29.1s	78.7%
	$M = 4$	29.7	61.9	48.3	37.9s	72.2%
	$M = 5$	31.9	64.3	51.7	47.2s	65.4%
	$M = 6$	31.8	64.3	53.1	56.9s	58.3%

$M = 5$ in the experiments here. A possible reason for the slightly worse detection performance of segment-based DTW is the restriction of the matched utterance fragments to the boundary of basic segments. In other words, frame-based DTW allows the matched fragments in the test utterance to start and end at any frame, while segment-based DTW only allows the matched fragments to start and end at the boundaries of basic segments. We can make the basic segments finer by choosing a smaller β , which gave better detection performance but the CPU time increased rapidly. This is only the first attempt on spoken term detection using segment-based DTW. The preliminary experimental results indicates that there is still much room for further improvements.

5. Conclusions

In this paper we propose segment-based DTW for unsupervised spoken term detection. We employ HAC with minimum sum of squared error objective [9] and proposed a stopping criterion to generate basic segments from speech signals. We then formulate the frame-based DTW with a new rate distortion constraint and the segment-based DTW given two sequences of basic segments. Although the detection performance degraded slightly in the preliminary experiments, the efficiency was improved significantly and much room is left for future work.

6. References

- [1] NIST. The spoken term detection (STD) 2006 evaluation plan, 10th ed. [Online]. Available: <http://www.nist.gov/speech/tests/std>
- [2] D. R. H. Miller *et al.*, “Rapid and accurate spoken term detection,” in *Interspeech*, 2007.
- [3] W. Shen, C. M. White, and T. J. Hazen, “A comparison of query-by-example methods for spoken term detection,” in *Interspeech*, 2009.
- [4] J. Mamou, B. Ramabhadran, and O. Siohan, “Vocabulary independent spoken term detection,” in *Proc. ACM-SIGIR*, 2007.
- [5] R. Wallace, R. Vogt, and S. Sridharan, “A phonetic search approach to the 2006 NIST spoken term detection evaluation,” in *Interspeech*, 2007.
- [6] M. Saraclar and R. Sproat, “Lattice-based search for spoken utterance retrieval,” in *HLT-NAACL*, 2004.
- [7] Y. Zhang and J. R. Glass, “Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams,” in *ASRU*, 2009.
- [8] T. J. Hazen, W. Shen, and C. White, “Query-by-example spoken term detection using phonetic posteriorgram templates,” in *ASRU*, 2009.
- [9] Y. Qiao, N. Shimomura, and N. Minematsu, “Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons,” in *ICASSP*, 2008.
- [10] A. S. Park and J. R. Glass, “Unsupervised pattern discovery in speech,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, 2008.