Unsupervised Query-by-Example Spoken Term Detection using Bag of Acoustic Words and Non-Segmental Dynamic Time Warping

Basil George, Abhijeet Saxena, Gautam Mantena, Kishore Prahallad and B. Yegnanarayana

Speech and Vision Lab, International Institute of Information Technology, Hyderabad,
{basil.george, abhijeet.saxena, gautam.mantena}@research.iiit.ac.in,
{kishore, yegna}@iiit.ac.in

Abstract

The paper proposes an unsupervised framework to address the problem of spotting spoken terms in large speech databases. A two-stage retrieval mechanism is used to perform spoken term detection. A very efficient Bag of Acoustic Words (BoAW) index is created for quick retrieval of relevant documents. Using an N-gram approach, the optimum choice of acoustic dictionary that best describes the document is obtained. Once a quick reduction in search space is achieved in the first phase, the results are fed to the second stage of the retrieval engine. Here, a computationally optimised variant of dynamic programming, called Non-Segmental Dynamic Time Warping (NS-DTW), is used to further prune the results. All the experiments are conducted on MediaEval 2012 dataset. Performance is evaluated at the output of each stage, and the optimum parameters are obtained. We show that the cascade of these two stages helps in reducing the probable search space, which translates to higher search speeds, while ensuring comparable performance. The significance of the indexing framework is proved by its comparison against a random selection system.

1. Introduction

Systems that search through spoken content have become very popular since the massive spread of multimedia content such as broadcat news, audio books, classroom lectures etc. Query-by-example (QbE) spoken term detection (STD) is a speech search framework in which spoken queries are used to retrieve matching portions from a speech database. While LVCSR-based systems have shown good performance for STD tasks, their resource intensive nature in terms of the requirement of transcribed data for training and their problem of out-of-vocabulary (OOV) terms make their use quite limited in many contexts. Though some methods like making the system vocabulary independent, sub-word unit modeling of OOV terms, phonetic search frameworks etc. have been proposed to address the OOV problem, it continues to be a challenging task [1, 2, 3, 4].

2. Related Work

As an alternative to LVCSR-based systems, template matching based methods have been explored in recent years for QbE STD [5, 6, 7, 8, 9]. In these methods, audio data is stored as templates that are generated by acoustic-phonetic models. When a spoken query is presented to the system, its template is generated, which is then searched in the database, typically by using a variant of the Dynamic Time Warping (DTW) algorithm. Posteriorgrams, which is a representation of speech as posterior probability vectors, have been widely used as templates recently [5, 6, 8, 10].

But the absence of efficient indexing techniques makes posteriorgram-based systems not scalable for practical use, as the entire database needs to be searched in a linear fashion irrespective of the query length. In [11], a Bag of Acoustic Words (BoAW) appraoch to QbE STD has been proposed. But the need for segmenting the documents before indexing, and the document segments to match the length of query segments for better results, may cause the system unscalable, especially when larger queries are introduced. This paper avoids these requirements, thus making the system scalable for large speech databases and for any length of the query. Robust indexing and retrieval techniques, along with a very efficient Non-Segmental DTW algorithm, have been used which makes it a very practical system.

3. Indexing and Retrieval of Bag of Acoustic Words

The BoAW model used in this work is similar to the Bag of Words (BoW) model widely employed in text retrieval systems. The vocabulary used to describe a document is obtained using Gaussian Mixture Modeling (GMM) of MFCC speech features. The number of Gaussian distributions (K) is the size of the acoustic dictionary used to transcribe the document.

3.1. N-gram approach

This acoustic vocabulary obtained through GMM modeling is then used to quantize the extracted features from documents by choosing the clusters with the highest posterior probabilities. This may be termed as a uni-gram appraoch. After extracting such uni-gram words, a word-document co-occurrence matrix is generated which describes a histogram of acoustic words in individual spoken documents. The histogram is the frequency count of the quantized acoustic features $[f_1, f_2, ...f_i, ...f_K]$, where f_i is the number of occurrences of the i^{th} cluster or acoustic word in the spoken document and K is the vocabulary size. Figure 1 shows the posteriorgram and histogram of a portion of a document using a uni-gram dictionary.

The histogram representation destroys the sequence information present in speech documents. To model the dynamic information embedded in acoustic words, an N-gram based approach is used, which describes partial dynamics of acoustic words by considering N consecutive words [12]. In this work, we evaluate the performance of the system by using bi-gram ($N = 2$) and tri-gram ($N = 3$) dictionaries, in addiditon to the uni-gram dictionary.

The bi-gram dictionary ($\mathcal{W}_b$) is created by combining the acoustic words which form the original uni-gram dictionary ($\mathcal{W}_u$). The i-th word in the new dictionary is defined as fol-

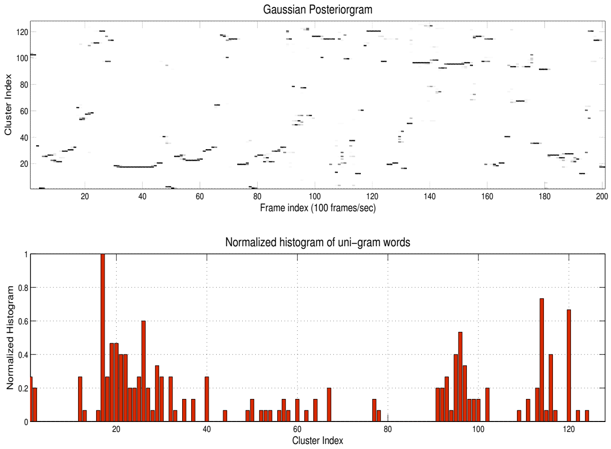

Figure 1: *(a) Gaussian posteriorgram and (b) normalized BoAW histogram of a speech segment with K=128, using uni-gram dictionary.*

lows:

$$w_b{}^i = \{(w_u{}^n, w_u{}^m) | w_u{}^n, w_u{}^m \in \mathcal{W}_u\}, 0 \le i < K^2 \quad (1)$$

where

$$n = \lfloor i/K \rfloor \quad (2)$$

$$m = mod(i/K) \quad (3)$$

where $\lfloor . \rfloor$ and $mod(.)$ represent the maximum integer that does not exceed the value of the division and the modulus of the division respectively. Similarly, tri-gram dictionary is created by combining words from the uni-gram and the bi-gram dictionaries. Note that the size of the bi-gram dictionary is K^2 and that of the tri-gram dictionary is K^3.

Once the dictionaries are created, all the spoken documents are described using uni-gram, bi-gram and tri-gram acoustic words. This N-gram approach helps to describe the local context, which is otherwise not possible in the BoW approach. Experiments are conducted to choose the optimum combination of dictionaries which gives the best performance. For simplicity, we choose uni-gram dictionary size to be $K = 128$. Hence, bi-gram dictionary size is $K^2 = 16,384$ and tri-gram dictionary size is $K^3 = 2,097,152$.

3.2. N-gram Indexing and Retrieval

The BoAW histogram tokens are stored in a very efficient inverted indexing data structure. An open-source information retrieval library, called Lucene [13], is used for indexing and retrieving the documents. Term frequencies (tf) and inverse document frequencies (idf) are used to score the documents when a query is presented to the system. A combination of Boolean Model (BM) and Vector Space Model (VSM) of information retrieval is used in the retrieval engine. In the Boolean Model, a document is retrieved if at least one of the query tokens is present in an indexed document. Once all the relevant documents are retrieved, the Vector Space Model is used to rank them in the order of relevance. The practical scoring formula used to compute similarity between a query vector q and a document vector d is given by Eq. 4.

$$S(q,d) = \sum_{t \in q : t \in d} (tf(t,d) * idf(t,D)^2 * Q_{boost}(t,q) * I(t,d)) \\ * \; C(q,d) * Q_{norm}(q) \quad (4)$$

where t denotes a term/token, $tf(t,d)$ is the term frequency of t in document d and $idf(t,D)$ is the inverse document frequency which denotes how rare the term is in the entire set of documents, D. They are given by:

$$tf(t,d) = \sqrt{f(t,d)} \quad (5)$$

and

$$idf(t) = 1 + log(\frac{N}{|d \in D : t \in d| + 1}) \quad (6)$$

where $f(t,d)$ is the frequency of t in d, N is the total number of documents in the corpus and $|d \in D : t \in d|$ is the number of documents in D where t appears. $Q_{boost}(t,q)$ is the search time boost factor, associated with the term t in query q, used to denote if some terms are more important than others. $I(t,d)$ encapsulates indexing time boost and document length normalization factors. $C(q,d)$ denotes the overlap of query within a document, so that the larger the overlap, the better the score. $Q_{norm}(q)$ is the query normalization factor used to make scores across different queries comparable. In this work, all the terms in a document and a query are assumed to be of equal significance, and hence, the indexing time and search time boost factors are both kept as 1. A detailed description of this scoring formula can be found in [13].

4. QbE-STD using Non-Segmental DTW

Once the first stage of retrieval using BoAW is done, QbE-STD is performed on the reduced search space using a variant of DTW, referred to as non-segmental DTW (NS-DTW) [14], to get the exact location of the query within the documents. This very fast DTW algorithm is performed on the Gaussian posteriorgrams of the reference and the query, which were computed earlier. The distance measure between a query vector $\mathbf{q_i}$ and a reference vector $\mathbf{u_j}$ is given by:

$$d(i,j) = -log\left(\frac{\mathbf{q_i}}{||\mathbf{q_i}||} \cdot \frac{\mathbf{u_j}}{||\mathbf{u_j}||}\right) \quad (7)$$

where $\mathcal{Q} = \{\mathbf{q_1}, \mathbf{q_2}, \ldots, \mathbf{q_i}, \ldots, \mathbf{q_n}\}$ is the query posteriorgram containing n feature vectors and $\mathcal{R} = \{\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_j}, \ldots, \mathbf{u_m}\}$ is the reference posteriorgram containing m feature vectors. The term *search hit* is defined as the region in the reference $\mathcal{R}$ that is likely to contain the query $\mathcal{Q}$. The query can start from any point in the reference. Initially, $S(1,j) = d(1,j)$, where $d(1,j)$ is the distance measure. The entries in the rest of the similarity matrix for NS-DTW is given by Eq. (8).

$$S(i,j) = min \begin{Bmatrix} \frac{d(i,j) + S(i-1,j-2)}{T(i-1,j-2)+1} \\ \frac{d(i,j) + S(i-1,j-1)}{T(i-1,j-1)+2} \\ \frac{d(i,j) + S(i-1,j)}{T(i-1,j)+1} \end{Bmatrix}, \quad (8)$$

where T is called the transition matrix. $T(i,j)$ represents the number of transitions required to reach i,j from a start

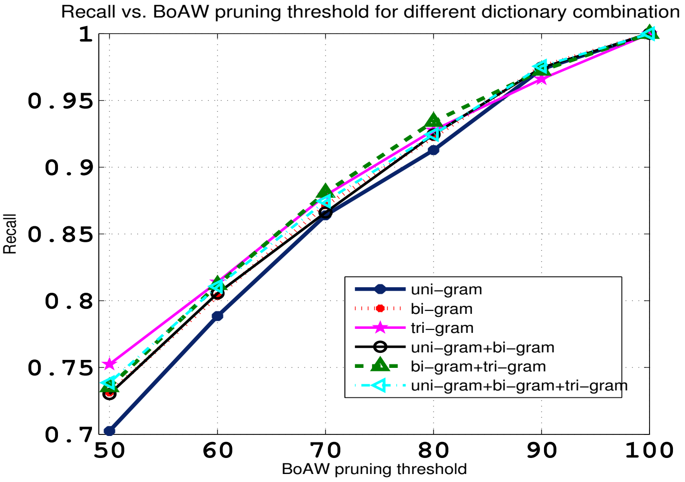

Figure 2: *Recall scores as a function of BoAW pruning threshold (δ) for different dictionary combinations.*

point. The location of the query within the reference is obtained by tracing back the path from the reference index that contains the best alignment score, i. e., from the end point of the *search hit* as given by $j = \min_{j}\{S(n,j)\}$ for $j = 1, 2, ..., m$. This algorithm, with a computational upper bound of $O(mn)$, helps in a quick search of the database, as compared to other existing template matching techniques.

5. Database

The experiments are conducted on MediaEval 2012 development (dev) and evaluation (eval) data sets, which consist of audio recorded via telephone in four African languages. The statistics of the audio data are shown in Table 1.

All the evaluations were performed using 2006 NIST evaluation criteria [15] and the corresponding maximum term weighted values (MTWV) are reported [16]. In addition to this, precision metrics like P@N and mean average precision (MAP) are also reported.

Table 1: Statistics of MediaEval 2012 data.

Data	#Utterances	Total(min)	Average(sec)
dev reference	1580	221.863	8.42
dev query	100	2.372	1.42
eval reference	1660	232.541	8.40
eval query	100	2.537	1.52

6. Experiments and Results

The experiments are conducted first on MediaEval development set to obtain the optimum system parameters, and then on the evaluation set with the new parameters. The first part of the experiments deals with obtaining the vocabulary of the BoAW system which is best suited for the task. The performance of this stage is evaluated using the rate of recall achieved for different pruning thresholds. Here, a trade-off between a reasonably high rate of recall and a good percentage reduction of the search space is made to obtain the optimum parameters.

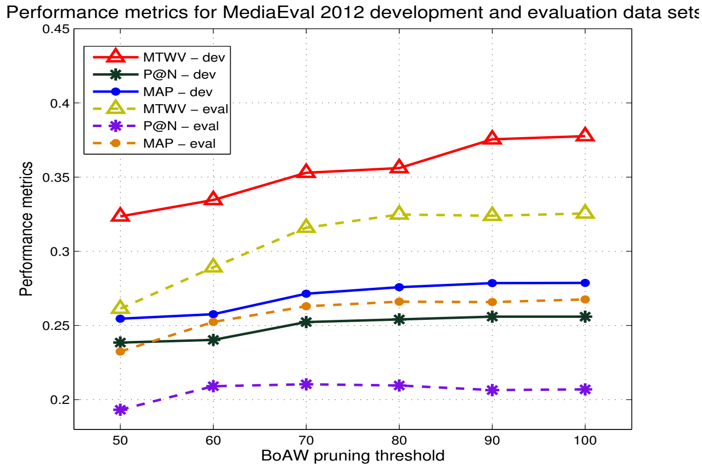

Figure 3: *Performance metrics for development and evaluation sets as a function of BoAW pruning threshold (δ).*

6.1. First stage: BoAW-based Inverted Indexing and Retrieval

First, features are computed on all the reference and query files by using the standard 39 dimensional MFCC features with frame size of $20ms$ amd frame shift of $10ms$. 128 dimensional posterior probability vectors are obtained for each feature vector using a trained GMM model. Vector quantization is performed and a Gaussian cluster label is assigned to each frame of the data. Then, uni-gram, bi-gram and tri-gram words are formed to describe the documents using the dictionaries that had been created when the GMM was trained. Uni-gram, bi-gram and tri-gram histograms for each of the reference and query files are created and are normalized. Thus, each document is described using three different bags of acoustic words and experiments are conducted to obtain the dictionary combination which gives the best recall for sufficient reduction in the search space.

Table 2: Recall scores for different BoAW pruning thresholds(δ) using bi-gram+tri-gram as dictionary.

δ	Recall:Dev.set	Recall:Eval.set
100	1	1
90	0.9724	0.9572
80	0.9341	0.9195
70	0.8810	0.8415
60	0.8119	0.7774
50	0.7354	0.6918

Figure 2 shows recall scores for different BoAW pruning thresholds and various dictionary combinations. If a bag or reference document containing at least one relevant result is returned by the BoAW retrieval system, then it is considered as a successful recall. BoAW pruning threshold (δ) determines the percentage of the top results retained for further search in the second stage of the STD system using NS-DTW algorithm. For example, if the pruning threshold is kept at 70, the new search space becomes the top 70% of results returned by the BoAW retrieval engine. From the figure, we learn that all the relevant results are retained ($recall = 1$) when all the entries returned by the system are preserved ($\delta = 100$), no matter what dictionary combination is used. This is equivalent to an absence in indexing, and hence can be used to compare the indexing performance. As we keep reducing the pruning threshold, more and more results get discarded as irrelevant entries, and hence, we see a gradual drop in the recall scores. This drop indicates that the description of documents using GMM labels is not ideal and hence, some of the relevant documents get very low ranks during BoAW scoring process, thus getting discarded as the pruning threshold drops. We trade this weakness for the huge gain we obtain in the time taken for search. Each query takes in the order of only a few milli-seconds to search the entire database. This very fast search capability, which translates to a very quick reduction in the search space, is the most important feature of this BoAW-based indexing approach to spoken term detection.

When the performance of different dictionaries is compared, it is observed that the document descriptions using bi-gram and tri-gram dictionaries perform better than the uni-gram dictionary. This is because bi-gram and tri-gram dictionaries can model the partial temporal characteristics of speech better than uni-gram descriptiors. Further, uni-gram + bi-gram + tri-gram combination of descriptors performs slightly inferior to bi-gram + tri-gram descriptors, thus suggesting that a model which does not incorporate the sequential nature of speech is better to be avoided while representing spoken documents, even in a bag of words paradigm. Thus, we use only the bi-gram + tri-gram dictionary combination for all our further experiments. Table 2 gives the recall scores for development and evaluations sets using this dictionary combination for different BoAW pruning thresholds (δ).

6.2. Second stage: NS-DTW

At the second stage of retrieval, a much more stringent temporal pattern matching algorithm is used to prune the results from the new reduced search space. The NS-DTW algorithm, as explained in section 4, is applied to the posteriors of pruned BoAW retrieval results. It returns precise locations within each reference file where hypothesized hits are detected. Mean Term Weighted Value (MTWV), the metric used in MediaEval 2012 evaluation, is used to compare the performance of the algorithm for different pruning thresholds. In addition to this, two other metrics are also used: P@N: average precision of the top N results, where N is the number of occurrences of each query in the database; and MAP: mean average precision, which is the mean of the precision scores after each query hit is retrieved.

Table 3: Comparison of MTWV scores on development set after BoAW indexing and random selection, both followed by NS-DTW .

δ	MTWV (BoAW)	MTWV (random)
90	0.3755	0.1992
80	0.3561	0.1588
70	0.3529	0.1916
60	0.3346	0.1314
50	0.3235	0.1104

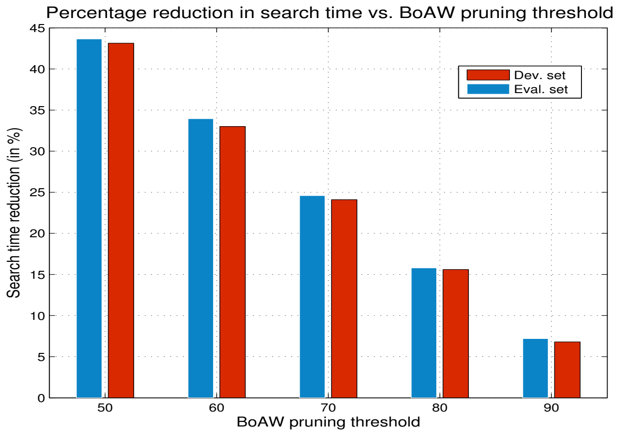

Figure 4: *Percentage reduction in search time for different BoAW pruning thresholds (δ) as compared to $\delta = 100$.*

Figure 3 shows the performance metrics for development and evaluation data sets as a function of δ. From the figure, it can be observed that MTWV scores drop slightly when δ is decreased. This is because some of the relevant results were removed by the first stage of indexing, as shown by the recall scores in table 2. Nevertheless, the drop in MTWV score is very less, particularly in the evaluation set, even when δ is reduced from 100 to 70. If the precision scores are analyzed, it can be seen that they remain largely the same when δ is dropped. This is very significant because precision scores tell how closer to the best/top result are all the relevant results returned by the system. The steady nature of the P@N curve shows that all the relevant entries are indeed preserved in the top N results, even when only 50% of the original database was used in the search. This translates to a reduction in search time as shown in figure 4. To indeed establish the relevance of the BoAW indexing stage, we compare its performance with a system in which random acoustic word bags are picked, instead of that returned after indexing, and then performing NS-DTW. The comparison provided in table 3 shows that the BoAW indexing engine, followed by the NS-DTW retrieval algorithm, consistently and significantly outperforms a system of random selection. This proves the efficiency of the indexing engine in filtering out irrelevant entries from the database, with practically no additional time overhead (of the order of only a few milli-seconds).

7. Conclusion

This paper proposes a very efficient two-stage approach to address the problem of unsupervised QbE STD: an efficient Bag of Acoustic Words (BoAW)-based indexing and retrieval stage, followed by a robust variant of DTW, called Non-Segmental DTW (NS-DTW). The data is indexed using a combination of N-gram tokens obtained from the Gaussian posteriorgrams. Using a very efficient scoring algorithm and by choosing an appropriate value for the pruning threshold (δ), the search space is reduced while preserving most of the relevant documents, for the next stage of search. In the second stage, the NS-DTW algorithm, with a computational upper-bound of $O(mn)$, retrieves the most relevant regions within documents belonging to this reduced search space. The entire search process is quickened with this two-pronged approach. Finally, the significance of the BoAW indexing stage is evaluated by comparing it with a system of random selection. The large difference in the performance proves the efficiency of the indexing stage in quickly discarding irrelevant documents.

8. References

[1] J. Mamou, B. Ramabhadran, and O. Siohan, “Vocabulary independent spoken term detection,” in *SIGIR*, 2007, pp. 615–622.

[2] I. Szöke, L. Burget, J. Cernocký, and M. Fapso, “Sub-word modeling of out of vocabulary words in spoken term detection,” in *SLT*, 2008, pp. 273–276.

[3] C. Parada, A. Sethy, and B. Ramabhadran, “Query-by-example spoken term detection for oov terms,” in *ASRU*, 2009, pp. 404–409.

[4] K. Ng, “Subword-based approaches for spoken document retrieval,” Ph.D. dissertation, Massachusetts Institute of Technology, 2000.

[5] T. Hazen, W. Shen, and C. White, “Query-by-example spoken term detection using phonetic posteriorgram templates,” in *ASRU*, 2009, pp. 421–426.

[6] Y. Zhang and J. R. Glass, “Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams,” in *ASRU*, 2009, pp. 398–403.

[7] C. an Chan and L.-S. Lee, “Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping,” in *INTERSPEECH*, 2010, pp. 693–696.

[8] V. Gupta, J. Ajmera, A. Kumar, and A. Verma, “A language independent approach to audio search,” in *INTERSPEECH*, 2011, pp. 1125–1128.

[9] A. Muscariello, G. Gravier, and F. Bimbot, “Zero-resource audio-only spoken term detection based on a combination of template matching techniques,” in *INTERSPEECH*, 2011, pp. 921–924.

[10] G. Aradilla, H. Bourlard, and M. Magimai-Doss, “Posterior features applied to speech recognition tasks with user-defined vocabulary,” in *ICASSP*, 2009, pp. 3809–3812.

[11] B. George and B. Yegnanarayana, “Unsupervised query-by-example spoken term detection using segment-based bag of acoustic words,” in *ICASSP*, 2014, pp. 7183–7187.

[12] S. Kim, S. Sundaram, P. G. Georgiou, and S. Narayanan, “An n-gram model for unstructured audio signals toward information retrieval,” in *MMSP*, 2010, pp. 477–480.

[13] “Lucene,” http://lucene.apache.org.

[14] G. V. Mantena, S. Achanta, and K. Prahallad, “Query-by-example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping,” *IEEE/ACM Transactions on Audio, Speech & Language Processing*, vol. 22, no. 5, pp. 944–953, 2014.

[15] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddington, “Results of the 2006 spoken term detection evaluation,” in *Proc. of Workshop on Searching Spontaneous Conversational Speech*, 2007, pp. 45–50.

[16] F. Metze, E. Barnard, M. H. Davel, C. J. V. Heerden, X. Anguera, G. Gravier, and N. Rajput, “The spoken web search task,” in *MediaEval*, 2012.