ELiRF at MediaEval 2013: Spoken Web Search Task

Jon A. Gómez, Lluís-F. Hurtado, Marcos Calvo, Emilio Sanchis
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camí de Vera s/n, 46020, València, Spain
{jon, lhurtado, mcalvo, esanchis}@dsic.upv.es

ABSTRACT

In this paper, we present the systems that the Natural Language Engineering and Pattern Recognition group (ELiRF) has submitted to the MediaEval 2013 Spoken Web Search task. All of them are based on a Subsequence Dynamic Time Warping algorithm and are *zero-resources systems*.

1. INTRODUCTION

In this paper, we present the systems that we have sumitted to the MediaEval 2013 Spoken Web Search task [2]. This task can be placed in the framework of Query-by-Example Spoken Term Detection (QbE-STD) tasks, where a set of documents and queries are provided, and the goal of the task is to find all the occurrences of each query within each document in the collection. In this particular case, a variety of languages and acoustic conditions are represented, but no information about them is provided to the participants.

All the systems we have submitted to this MediaEval 2013 Evaluation are based on a Subsequence Dynamic Time Warping (S-DTW) algorithm [1], but using different distances, sets of possible movements, and feature vectors. Also, all our systems are *zero-resources systems*, that is, they do not use any external information, but just the one provided by the task.

2. DESCRIPTION OF THE SYSTEMS

For this task, we have submitted four different systems, all of them based on the S-DTW algorithm. S-DTW is a Dynamic Programming (DP) algorithm which aim is to find multiple local alignments of two input sequences of *objects* using a set of allowed movements, but allowing one of the sequences to *start* at any position of the other. Equation 1 shows the generic formulation of the S-DTW algorithm.

$$M(i,j) = \begin{cases} +\infty & i < 0 \\ +\infty & j < 0 \\ 0 & j = 0 \\ \min\limits_{\forall(x,y)\in S} M(i-x, j-y) + D(A(i), B(j)) & j \geq 1 \end{cases} \quad (1)$$

where M is the DP matrix; S is the set of allowed movements, represented as pairs (x, y) of horizontal and vertical increments; $A(i)$, $B(j)$ are the objects representing the positions i and j of their respective sequences; and D is a function that computes some distance or dissimilarity between two objects.

In this task the sequences of objects to be aligned are the sequences of feature vectors obtained from the audio files corresponding to the documents and the queries. This approach allows us to find the best alignment of the query in each document taking as the starting point every frame of the document. For this work we have used the cosine distance in all our systems, since it provided the best results for the development set.

Each of the computed alignments should be considered a candidate detection. Hence, this strategy provides a too large number of candidates. This way, it is necessary to find a criterion to find the set of definitive detections among the elements of this set.

Another common step for all our systems is that, as part of the preprocessing, we deleted the leading and trailing silences of the queries by using a Voice Activity Detection strategy based on a Smith trigger. This led our systems to a better performance.

Thus, our systems differ basically on three different aspects: how the feature vectors are obtained, how to determine which of the candidate detections are considered as definitive, and which are the allowed movements in the Dynamic Programming algorithm.

2.1 System 1

In this system, the acoustic signal is parametrized using the energy, the first twelve cepstral coeficients and their first and second derivatives, using a sampling period of 10 ms. Thus, we represent each frame as a 39-dimensional vector. Then, we perform the S-DTW step, using a particular set of movements: {(1,2), (1,1), (2,1)}. Also, in each step of the S-DTW algorithm, we have kept and maximized the accumulated distance normalized by the number of operations carried out until that point. The set of candidate detections are all the hypotheses that arrived to any cell corresponding to the last frame of the query in the DP matrix. Furthermore, the set of movements used guarantees the size of any detection will be between 0.5 and 2 times the size of the query. These candidates are filtered using Algorithm 1. The idea of this algorithm is to find all the local minima that do not overlap any other local minimum with a better score, and then fix a threshold according to a linear combination of the average and standard deviation of the scores of the "cleaned" set of local minima (the parameter λ of this linear combination is empirically adjusted). Also, a maximum number of filtered detections for each query d is allowed. In this system, we have adjusted the parameters in order to

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

obtain just a few definitive detections per query.

Algorithm 1 Algorithm to filter a list of candidate detections

```
Require: A list of candidate detections $CD$,
    a maximum number of filtered detections $d$,
    a coefficient $\lambda$
Ensure: A list of filtered detections $FD$
$SCD$ = sort the hypothesis in $CD$ by their score
$FD2$ = empty list
while $SCD$ is not empty do
    $h$ = first element of $SCD$
    Move $h$ to $FD2$
    Delete from $SCD$ all the detections $h'$ such that
        timespan($h'$)$\cap$timespan($h$) $\neq \emptyset$
end while
$t = \text{avg} + \lambda \cdot \text{sd}$, where avg and sd represent the average
    and the standard deviation of the elements in FD2
$FD$ = first $d$ elements of $FD2$ with a score $\geq t$
return $FD$
```

2.2 System 2

This system is very similar to System 1, but the thresholds were adjusted in a less restrictive way. The number of hypotheses provided by this system is much larger than for System 1.

2.3 System 3

This system uses the same parametrization as Systems 1 and 2. However, the allowed movements for the S-DTW are {(0,1), (1,0), (1,1)}. Also the algorithm to filter the candidate detections is a bit different (see Algorithm 2). In this algorithm the condition for local minima not to be pruned is that: (i) they have a value larger than a threshold and (ii) there is not any other detection with a better score within a window of 2 seconds. Finally, at most n occurrences per query and k detections per document are allowed.

Algorithm 2 Another way of filtering a list of candidate detections

```
Require: A list of candidate detections $CD$,
    a maximum number of occurrences per query $n$,
    a maximum number of detections per document $k$
Ensure: A list of filtered detections $FD$
$SCD$ = empty list
for all Query $q$ do
    for all Document $d$ do
        m = minimum score of a detection of $q$ within $d$
        M = maximum score of a detection of $q$ within $d$
        $t = m + 0.1(M - m)$
        Add to $SCD$ all the hypotheses from $CD$ with a
            score larger than $t$ and that do not overlap a better
            detection within a window of 2 seconds.
    end for
end for
$FDP$ = For each query, keep the at most $n$ best occurrences in $SCD$
$FD$ = For each document, keep the at most $k$ best detections in $FDP$
return $FD$
```

2.4 System 4

This system is similar to System 3, but the way of obtaining the feature vectors varies. The features are here obtained by using a Dissimilarity Space. 300 frames are selected from the development set applying the Katsavounidis criterion with the cosine distance as metric [3]. Then each frame is moved into the dissimilarity space, where each component of the new feature vectors is computed as the distance from the sample to each one of the 300 taken as references. Thus, in this system the feature vectors have 300 dimensions. All the frames from both the documents and the queries are converted to this Dissimilarity Space, and the S-DTW is performed using these vectors.

3. EXPERIMENTS AND RESULTS

For this MediaEval 2013 Spoken Web Search Evaluation, we submitted one run for each of the four systems described above. The results we obtained are shown in Tables 1 and 2, where P stands for Precision and R means Recall. Table 2 also shows the Real Time factor (RT) obtained for the test set. Its value for the development set is very similar.

Table 1: Results obtained for the development set.

System	MTWV	ATWV	P(%)	R(%)	C_{nxe}
Sys. 1	0.1699	0.1697	3.47	15.69	2.45
Sys. 2	0.1296	0.1291	2.21	16.71	3.91
Sys. 3	0.1480	0.1478	3.18	14.37	1.03
Sys. 4	0.1463	0.1461	2.55	15.76	1.00

Table 2: Results obtained for the test set.

Sys.	MTWV	ATWV	P(%)	R(%)	C_{nxe}	RT
S. 1	0.1593	0.1591	3.29	14.89	2.53	$3\cdot10^{-3}$
S. 2	0.1016	0.1016	1.99	12.44	4.83	$3\cdot10^{-3}$
S. 3	0.1481	0.1475	3.03	13.66	1.03	$5\cdot10^{-4}$
S. 4	0.1462	0.1457	2.47	15.08	1.00	$2\cdot10^{-3}$

All the software of the systems presented here was completely developed in our research group. Also, all these systems were run on a standard PC with an i7 processor and 32 GB of RAM, using 8 threads. The memory peaks for systems 1 and 2 were around 12 GB, and for systems 3 and 4 were around 1 GB.

4. ACKNOWLEDGEMENTS

Work funded by the Spanish Government and the E.U. under contract TIN2011-28169-C05 and FPU Grant AP2010-4193.

5. REFERENCES

[1] X. Anguera and M. Ferrarons. Memory efficient subsequence DTW for Query-by-Example spoken term detection. In *2013 IEEE International Conference on Multimedia and Expo.*

[2] X. Anguera, F. Metze, A. Buzo, I. Szoke, and L. J. Rodriguez-Fuentes. The Spoken Web Search Task. In *MediaEval 2013 Workshop*, 18-19 October 2013.

[3] I. Katsavounidis, C.-C. Jay Kuo, and Z. Zhang. A new initialization technique for generalized Lloyd iteration. *Signal Processing Letters, IEEE*, 1(10):144–146, 1994.