ELiRF at MediaEval 2014: Query by Example Search on Speech Task (QUESST)

Marcos Calvo, Mayte Giménez, Lluís-F. Hurtado, Emilio Sanchis, Jon A. Gómez
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camí de Vera s/n, 46020, València, Spain
{mcalvo, mgimenez, lhurtado, esanchis, jon}@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 2014 Query by Example Search on Speech task. All of them are based on a Subsequence Dynamic Time Warping algorithm and do not use any other information from outside the task (*zero-resources systems*).

1. INTRODUCTION

In this paper, we present the systems that we have submitted to the MediaEval 2014 Query by Example Search on Speech task. The goal of the task is to identify the audio documents which match a spoken query. This match can be either exact (the same term both in the query and in the document), or with variations [2].

The two systems we have submitted are based on a Subsequence Dynamic Time Warping (S-DTW) algorithm [1]. However, the systems differ in the way the audio files are preprocessed, which makes the feature vectors to be different for each system. It is worth to note that this approach does not use any external information, which makes our systems *zero-resources systems*. In the following sections, we will explain the differences in how the feature vectors are computed for each system, the search algorithm, and the results obtained in this evaluation.

2. OVERVIEW OF THE SYSTEMS

Both of our systems used the same philosophy. First step was preprocessing all the audio files, both spoken documents and queries. This way we obtained a sequence of feature vectors as a representation of each audio file. Then, we took each possible pair (*document, query*) and run a S-DTW algorithm on them. This provided the bounds of a possible detection of the query within the document, and a score for this detection. Finally, a decision-making module established a threshold based on the scores of all the possible detections. This was necessary to provide detections with the highest confidences.

3. PARAMETRIZATION

We used Mel Frequency Cepstral Coefficients (MFCC) plus energy as the features for representing speech samples (audio documents and queries) as sequences of feature vectors. Each feature vector contains 39 values: the energy, the first 12 MFCCs plus the first and second time derivatives of the original 13 features.

We worked with two parametrizations. One is the standard feature extraction without applying any filter for noise reduction (labeled as **non-filtered**), and the other one consists on compensating the MFCC for improving the recognition on noisy speech (labeled as **Choi**) [3].

In both cases a preemphasis filter over the signal was applied with $H(z) = 1 - 0.95z^{-1}$, the subsampling frequency was $100Hz$, i.e. one feature vector every 10 ms, and a Hamming window of 25 ms was applied before computing the FFT. The first difference is found when computing the Mel-filterbank. Each filterbank output is computed as follows in the case of the non-filtered parametrization:

$$m_j = \log(Y_j)$$

where Y_j is the output magnitude of the j-th Mel-filterbank. In the case of using the approach proposed by Choi [3]:

$$m_j = \alpha_j \ \log\{1 + \beta_j \ max(Y_j - \hat{N}_j, \gamma_j \ Y_j)\}$$

where $\beta_j = 0.001$ and $\gamma_j = 0.4 \ \forall j$ in our implementation, $\hat{N}_j$ is the noise magnitude estimation of the j-th Mel-filterbank output, and

$$\alpha_j = \frac{\log(1 + \frac{Y_j}{\hat{N}_j})}{\sum_{k=1}^{M} \log(1 + \frac{Y_k}{\hat{N}_k})}$$

M is the total number of Mel-filters. α values are computed for each feature vector.

Next step in the parametrization is to compute the standard Discrete Cosine Transform to the Mel-filterbank. The first 12 MFCC are obtained. But in the case of the filtered parametrization a transformation of energy and each MFCC component is performed based on the Cumulative Distribution Mapping (CDM) technique [3], which is based on the use of histogram equalization originally developed for image processing [4]. Last step of parametrization was the computation of first and second time derivatives.

It is worth to note that most of queries contain leading and trailing silences. Therefore, we trimmed the sequence of feature vectors representing each query by means of a voice activity detection procedure, in order to help the search algorithm.

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

4. SEARCH ALGORITHM

Finding spoken queries within a set of audios is a complex task, hence we used a Dynamic Programming (DP) technique in order to face this problem. In particular, we used S-DTW, that is a DP technique for comparing two sequences of objects. In our case, one of the sequences corresponds to feature vectors of one of the audio documents, and the other one belongs to the query. Therefore, the S-DTW method finds multiple local alignments of the query within audio documents, by allowing it to *start* at any position of the audio document.

Equation 1 shows the generic formulation of S-DTW:

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

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-th and j-th of their respective sequences; and D is a function that computes some distance or dissimilarity between two objects.

In our implementation the set of allowed movements S is $\{(1,2),\ (1,1),\ (2,1)\}$. This set of movements guarantees that the size of any detection will be between 0.5 and 2 times the size of the query.

5. EXPERIMENTS AND RESULTS

We performed several preliminary experiments in order to find the best configuration for our systems.

We evaluated different distance functions and parametrizations. One of them was the Kullback-Leibler divergence on sequences of vectors of probabilities as representation of audio files. The probabilities were obtained by a GMM estimated by means of the EM algorithm with all the audio documents in the corpus. Different number of components in the GMM were tried. We also tried the cosine distance with the Mel-filterbank parametrization. However, we finally used cosine distance with the MFCC, since it provided the best results for the development set.

For this MediaEval 2014 Query by Example Search on Speech Evaluation, we submitted one run for both systems described above. The results we obtained are shown in Tables 1 and 2. The measure to be optimized for this Evaluation was the cross entropy score (Cnxe). However, other measures such as the Mean and the Actual Term Weighted Values (MTWV and ATWV, respectively) were considered as secondary metrics, as they are very widely used in this kind of tasks.

Results shown in both tables reveal a bad performance of our systems (a high value of Cnxe). Nevertheless, given the difference in the sources of the audio documents and audio queries, we expected a higher accuracy for our system that uses the Choi parametrization.

We run our own multi-thread implementation of S-DTW algorithm, using an standard PC with an i7 core and 16 GB of RAM using 8 threads on a Linux operating system. At the parametrization stage, we achieved an *indexing speed factor* of $1.26 \cdot 10^{-2}$, and our memory peak was around 0.25 GB. At the search stage, our *searching speed factor* was $2.34 \cdot 10^{-3}$ and we used around 16 GB of memory. Thus, our *processing load* for both systems was $3.40 \cdot 10^{-2}$.

Table 1: Results obtained for the development set.

System	Cnxe	MinCnxe	MTWV	ATWV
Choi	5.8940	0.9595	0.0692	0.0692
non-filtered	6.0905	0.9571	0.0767	0.0768

Table 2: Results obtained for the test set.

System	Cnxe	MinCnxe	MTWV	ATWV
Choi	5.5369	0.9667	0.0558	0.0557
non-filtered	5.9502	0.9648	0.0587	0.0587

6. CONCLUSIONS

In this paper, we have presented the systems we have submitted to the MediaEval 2014 Query by Example Search on Speech Evaluation, as well as the results obtained. This was a very challenging task in which both exact and varied occurrences of the queries within the documents had to be found. Despite of our preliminary attempts, our approach has been proven as not suitable for this task. One of the reasons is due to the nature of the S-DTW algorithm. Its use makes not possible to find occurrences of queries where a reordering of words is needed. However, we would like to point out that significant improvements were observed when trimmed queries were used for the development set.

As future work, we would like to improve our system in order to use it in tasks like QUESST, where swaps in the order of components of a query can happen. Facing this kind of word reorderings would be possible if a higher level of knowledge is used, e.g. sequences of phonemes instead of using only sequences of acoustic feature vectors. It is not possible to use words in a task where distinct languages may appear and no other source than audio files is provided.

7. ACKNOWLEDGMENTS

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

8. 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*. IEEE, 2013.

[2] X. Anguera, L. J. Rodriguez-Fuentes, I. Szöke, A. Buzo, and F. Metze. Query by Example Search on Speech at Mediaeval 2014. In *MediaEval 2014 Workshop*, 16-17 October 2013.

[3] E. H. C. Choi. On compensating the mel-frequency cepstral coefficients for noisy speech recognition. In *Proceedings of the 29th Australasian Computer Science Conference - Volume 48*, ACSC '06, pages 49–54, Darlinghurst, Australia, Australia, 2006. Australian Computer Society, Inc.

[4] J. C. Russ and R. P. Woods. The image processing handbook. *Journal of Computer Assisted Tomography*, 19(6):979–981, 1995.