# The Telefonica Research Spoken Web Search System for MediaEval 2013

Xavier Anguera
Telefonica Research
Barcelona , Spain
xanguera@tid.es

Miroslav Skácel
Speech@FIT
Brno, Czech Republic
xskace00@stud.fit.vutbr.cz

Volker Vorwerk
Altran
Barcelona , Spain
volker.vorwerk@altran.com

Jordi Luque
Telefonica Research
Barcelona , Spain
jls@tid.es

## ABSTRACT

In this paper we describe the system proposed by Telefonica research for the Spoken Web Search (SWS) task [3] within the Mediaeval 2013 evaluation. This is the third year we participate in the evaluation and this time we have submitted a system based on the recently proposed Information Retrieval-based Dynamic Time Warping (IR-DTW) Algorithm. This algorithm performs a pattern matching search at frame level similar to the DTW algorithm, but with advantages in memory usage and the possibility to with pre-index the search corpora and use fast retrieval techniques. Results obtained this year have been poorer than expected, most probably due to the use of a global voice activity detector that was not adequate to the varying nature of the different acoustic conditions in this year's search corpora.

## 1. INTRODUCTION

In this paper we present the Telefonica research system proposed for the Spoken Web Search (SWS) task within the Mediaeval 2013 evaluation [3]. The task of Query-by-Example Spoken Term Detection (QbE-STD), also coined as Spoken Audio Search (SAS), has gained quite a bit of interest in the latter years within the scientific community as it allows for information to be obtained from audio documents whose language and/or acoustic conditions are not matched with those for which plenty of resources are available, and thus systems based on supervised training techniques can not be usually employed.

To tackle this year's evaluation we have implemented a zero-resources (no external data is used) system based on a frame-based pattern matching approach using the recently proposed IR-DTW algorithm [2] . The IR-DTW algorithm is inspired on the subsequence-DTW algorithm [5] to which we add the option to pre-index the search corpus for faster retrieval and a dynamic programming algorithm inspired on information retrieval techniques which allows us to perform a time-warped matching with very limited memory requirements. Results for this year's evaluation are poorer than we expected, partly due to the use of a global voice activity detector that is not well coupled with the varying acoustic conditions of this year's test data.

## 2. THE SWS SYSTEM DESCRIPTION


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

Figure 1 shows the main components of the system we presented his year. We first perform feature extraction on the audio signal to obtain standard MFCC features (12 Cepstra+energy + $\Delta$ + $\Delta\Delta$) which are mean and variance normalized. Then, we obtain 64-dimensional posterior probabilities from these features using a modified background model described in Section 2.1. Optionally, the search corpus features can be indexed into a hierarchical tree structure to speedup later search, as described in [4]. Next, we label the audio into speech/non-speech frames using a global voice activity detector (Section 2.2) and eliminate the non-speech frames from the pattern matching steps. Next we apply the IR-DTW algorithm (Subsection 2.3 ) to find all matches for every given query sequence. Finally, the top 500 matches are processed using a more exhaustive local subsequence-DTW matching to obtain exact start-end points and their scores. Z-Norm is applied to the final set of results to make scores comparable across queries. In the case that we have multiple instances per query we use the standard system on each query and then fuse the results as described in Section 2.4.

### 2.1 Background Model Creation

The usage of a Gaussian Mixture Model (GMM) to obtain posterior probability features from the input features was introduced to QbE-STD by [7]. In [1] we proposed an alternative to the standard GMM model by training a background model that focuses on creating more discriminative Gaussian mixtures. A novelty we introduce this year for our system is inspired on last year's system submission from CUHK [6] where VTLN is applied to obtain a vocal tract normalized GMM model. In our system we initially train a background model as described in [1] by using all search corpus data. Then we estimate VTLN coefficients for each utterance in the corpus by using this model. The VTLN-normalized features are then used to train a new background model from scratch. The process is repeated 3-4 times until convergence.

### 2.2 Voice Activity Detection

It is well known that silence/non-speech regments are not suitable for frame-based pattern-based matching as they tend to generate many false alarms. For this reason, it is mandatory to detect such frames and mask them out in subsequent matching steps. To determine whether a frame is speech or non-speech we use two global GMM models trained on the whole search corpus for speech and non-speech. The initialization of the models is done by using the 10% least energy frames for the silence model, and the rest for the speech model, followed by several decoding and retraining

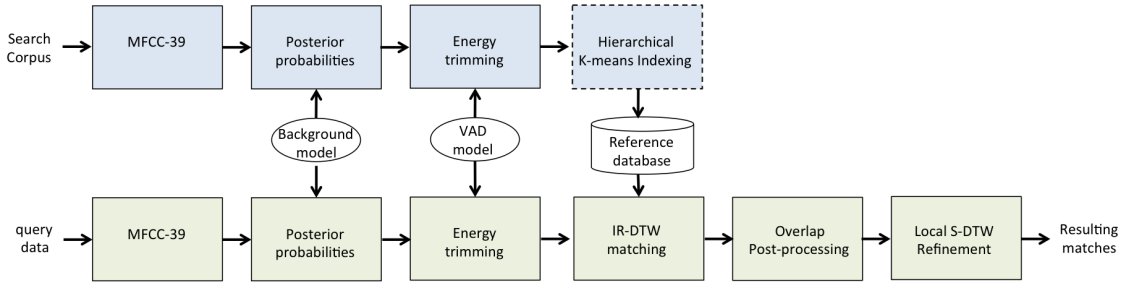


**Figure 1:** *General system blocks for the Telefonica system*

iterations. For the evaluation we used 2 Gaussians for non-speech and 16 Gaussians for speech. Assignment of an input frame to speech or non-speech is done through hypothesis testing with both models. A final morphological filter is run to smooth the speech and non-speech regions.

## 2.3 IR-DTW and Overlap Postprocessing

The IR-DTW algorithm used is an evolution of the algorithm presented in [2]. The IR-DTW algorithm performs a *sparse* dynamic programming pattern-matching algorithm only on the pairs of query-search corpus frames that exceed a certain similarity threshold, combining them to form possible sequences matching both query and corpus utterances. In addition, through a projection of the similarity matrix into a one-dimensional structure, the IR-DTW algorithm has much smaller memory requirement. Given all possible matching subsequences in the search corpora, many are in temporal overlap with each other. We filter out these overlapping paths in two steps: first we merge all those paths that have the same start time both in the query and in the search corpora and then we merge those matching paths that highly overlap (more than 50%) with each other.

## 2.4 Multiple Query Instances Fusion

In addition to the regular system, we implemented a straightforward fusion to take advantage of cases when multiple instances of a query are available. In such cases we run the system independently for each query and then merge together the z-normalized results for each of the instances. In the process we merge any overlapped results, keeping the one with highest score.

# 3. RESULTS

Table 1 shows the results obtained by our system on the dev and eval queries, both for the core set (only one instance of each query is available) and extended set. Results are lower than expected. After some preliminary testing we observed that the selection of a global VAD was a bad choice, as it is not able to do a good job in classifying speech/non-speech for the different acoustic conditions present in the database. In addition, we observe that we did not do a good job setting the optimum threshold on the ATWV score (neither with the dev or eval queries). In a positive tone, the extended queries results are consistently better (on MTWV) than the single queries, which means that our late fusion of results works.

To run the system we used standard desktop linux machines (approx. 2.5GHz processor speed) executing the program in single core. The amount of RAM memory used (discounting the storage of the database in memory) is around 110MB in average. The run-time factor is around 1e-3 RT.

**Table 1: Official Evaluation Results**

| Subset | Metric | dev queries | eval queries |
|---|---|---|---|
| core | MTWV | 0.1158 | 0.0925 |
| core | ATWV | 0.0961 | 0.0793 |
| extended | MTWV | 0.1303 | 0.1035 |
| extended | ATWV | 0.0845 | 0.0928 |

for the core set and 8.3e-4 RT for the extended set (core + extended queries) for the online search (feature extraction and background model computation are excluded from these measurements). These values are improved from last year's system thanks to a big rework we did of our system and the optimization of all algorithms. We still feel that current processing speeds should be increased for the system to be usable in a real-life implementation..

# 4. CONCLUSIONS AND FUTURE WORK

In this paper we presented the Telefonica system proposed for the SWS task within Mediaeval 2013. The system is based on frame-based pattern-matching using a novel DTW algorithm called IR-DTW. The results are far from expected, due in part to the use of a global VAD that does not adapt well to the varied acoustic conditions present in the search corpus. We intend to understand and solve these problems and to extend the IR-DTW matching algorithm to perform dynamic programming symbol-based matching over phoneme lattices in order to have a second set of results that can be later fused with the frame-based results.

# 5. REFERENCES

[1] X. Anguera. Speaker Independent Discriminant Feature Extraction for Acoustic Pattern-Matching. In *Proc. ICASSP*, 2012.

[2] X. Anguera. Information retrieval-based dynamic time warping. In *Proc. Interspeech*, Lyon, France, 2013.

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

[4] G. Mantena and X. Anguera. Speed Improvements to Information Retrieval-Based Dynamic Time Warping using Hierarchical k-Means Clustering. In *ICASSP*, 2013.

[5] M. Müller. Dynamic Time Warping, chapter 4. In *Information Retrieval for Music and Motion*, pages 69–84. Springer-Verlag, Berlin, Germany, 2007.

[6] H. Wang and T. Lee. CUHK System for the Spoken Web Search task at Mediaeval 2012. In *in Proc. Mediaeval workshop*, 2012.

[7] Y. Zhang and J. R. Glass. Unsupervised Spoken Keyword Spotting via Segmental DTW on Gaussian Posteriorgrams. In *Proc. ASRU*, pages 398–403, Merano, Italy, 2009.