Telefonica Research System for the Spoken Web Search task at Mediaeval 2012

Xavier Anguera
Telefonica Research
Edificio Telefonica - Diagonal 00
08019 Barcelona, Spain
xanguera@tid.es

ABSTRACT

In this paper we describe the systems presented by Telefonica Research to the Spoken Web Search task of the Mediaeval 2012 evaluation. This year we proposed two systems. The first one consists on a segmental DTW system, similar to the one presented in 2011, with a few improvements. The second system also uses a DTW-like approach but allowing for all reference files o be searched at once using an information retrieval approach.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.7 [**Artificial Intelligence**]: NAtural Language Processing

General Terms

Algorithms, Experimentation, Languages

Keywords

Dynamic Time Warping, Spoken Web Search, Low Resources

1. INTRODUCTION

The task of searching for speech queries within speech reference files is gaining interest in the scientific community as it does not require the knowledge a priori of the language or acoustic conditions of the data. This is specially relevant for languages with limited resources where not sufficient training data is available to build full-fletches speech recognition systems. Within the Spoken Web Search task (SWS) in the Mediaeval evaluation campaign for 2012 [2] systems are given a set of acoustic queries that have to be searched for within a corpus of audio composed of several African languages. No information about the transcription of the queries or reference data, nor the language spoken in which each query or reference file is given.

To tackle this task we propose two systems using a zero-resources approach derived from the well-known dynamic time warping (DTW) algorithm. Our main submission is composed of a variation of the DTW that allows us to search for a query within all reference data all at once, and to use standard information retrieval techniques to speedup the search process while still obtaining an accurate match (we will call it IRDTW from now on). Our secondary submission consists on the system we submitted for the 2011 evaluation with some small improvements (which we will call SDTW). Both systems use a common general framework we will describe next.

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy

2. GENERAL SWS FRAMEWORK

Figure 1 shows the main blocks that conform both systems we presented this year. First we extract MFCC-39 (13+13+13) features from the acoustic data in a standard manner (10ms scroll in 25ms window). Then we compute posterior probabilities from these frames using a posteriors background model we describe below. Then, after labeling the energy level of each frame we decide whether each frame is to be considered for matching. The matching is either using IRDTW or SDTW. Finally, we postprocess the results to eliminate overlapping results and we return the results and their scores.

2.1 Posteriors Background Model

Posterior probabilities have been successfully used in pattern matching for some time [3]. Several methods have been proposed to obtain the posterior probabilities. In our systems we use Gaussian posteriors obtained from a GMM model that has been trained on all available reference data (i.e. development and testing data). The GMM has been trained using a combination of EM and K-means iterations in order to maximize the discovery and separation of automatically discovered acoustic regions in the acoustic space. For more information on the model refer to [1].

2.2 Speech/silence labeling

One of the biggest enemies of pattern matching approaches is silence, as silence usually matches very well with silence, thus returning many false alarms unless it can be trimmed back. To eliminate silence from our input data (both queries and references) this year we trained a speech/silence classifier using GMM models in a non-supervised way. First, we gather the 10% of acoustic frames with lowest energy from our training data (the reference files in the development set). With these frames we train a one Gaussian silence model and with the rest we train a 4-Gaussian speech model. Then we iteratively assign each frame in the training set to the closest model and retrain the models. This usually increases the number of frames in the silence model. We stop after 20 iterations or when the difference in number of frames between two consecutive iterations is very small. We store the Gaussians in the speech model ordered by their mean en-

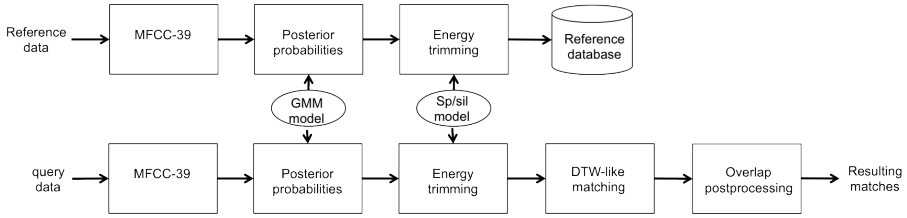

Figure 1: *General system blocks for SWS task*

ergy. In order to label the data using this model we no only assign each frame to the most likely model, but in the case of speech, we record which of the Gaussian mixtures is the closest one. In test we consider as silence (and therefore do not use for matching) any frame that is labelled as silence or is assigned to the lowest speech Gaussian.

2.3 Matching segments overlap detection

Common to both systems proposed for this year, the overlap detection module is used on all matching segments returned by the matching module to reduce the number of segments finally returned as matches. Last year we used a simple overlap detector to find when a segment from the query and the reference data were in overlap with any other detected segment. If the overlap was greater than 50% we merged both segments into one, considering the smallest of the start and the greater of the end positions. For this year we have improved the algorithm to consider the situations where a segment is fully embedded within the other segment both in query and reference. In addition, we observed that the merging criterion was not good enough, as this year an accurate alignment of the reference data is used and systems are expected to find the exact location of the query in such data. For this reason once we detect two overlapping paths we select the one with higher matching score and ignore the other one.

3. SUBMITTED SYSTEMS

In this section we describe the two matching algorithms we submitted to the evaluation.

3.1 Dynamic Time Warping using Information Retrieval techniques - IRDTW

This constitutes our primary submission this year. It consists on a rework of last year's DTW implementation to allow for the use of information retrieval techniques (like LSH, locality sensitive hashing) to speedup the matching between query and reference data. First, all reference data is loaded into memory at once, and the algorithm is then called for each query term. Matching is performed sequentially (looking at each frame in the query term) but considering only the best matching frames in the reference data according to their similarity to the query frames. In the current implementation the algorithm returns a list of possible matching segments (both in query and reference) that not necessarily cover all the query from start to end. To obtain an accurate score for the whole query we further perform a segmental-DTW between the start/end of the query and the start/end points of the found paths. In addition, we eliminate from all returned paths those that are in big overlap between each other (which happens usually, as more than one path can be always found for regions in the similarity matrix with high similarity).

Table 1: Official Evaluation Results

System	Metric	dev-dev	dev-eval	eval-dev	eval-eval
IRDTW	MTWV	0.390	0.314	0.498	0.342
	ATWV	0.386	0.304	0.422	0.330
SDTW	MTWV	0.374	0.300	0.472	0.311
	ATWV	0.364	0.292	0.399	0.294

3.2 Segmental Dynamic Time Warping - SDTW

This system is very similar to last year's submission. This year we only experimented (due to lack of time) with the real-valued input features, leaving for future work to test this year's data using the binary features, which obtained good results in last year's evaluation. The main differences of the system for this year have to do with the treatment of the speech/non-speech regions and the detection of overlapping matches, described above.

4. OFFICIAL RESULTS

Table 1 shows the results obtained by our systems. We can see how the IRDTW system obtained better results than the SDTW system we proposed last year, even though the IRDTW system does not take into account all distances between query and reference for the dynamic programming step.

5. CONCLUSIONS

This years participation has focused on the implementation and testing of an algorithm that allows us to scale, processing a large amount of reference data. Our primary algorithm is able to accomplish this. Although at the present time it is still not as fast as other proposals, we see no degradation of its accuracy in comparison to exact implementations (represented with our las year participation). In the near future we will implement speedups to the algorithm to use it with hundreds of hours of data. Also, we will work on improving the accuracy of results.

6. REFERENCES

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

[2] F. Metze, E. Barnard, X. Anguera, and G. Gravier. The Spoken Web Search Task. In *Proc. Mediaeval Workshop*, 2012.

[3] Y. Zhang and J. R. Glass. Unsupervised Spoken Keyword Spotting via Segmental DTW on Gaussian Posteriorgrams. In *Proc. ASRU*, 2009.