The TUM Cumulative DTW Approach for the Mediaeval 2012 Spoken Web Search Task

Cyril Joder, Felix Weninger, Martin Wöllmer
Institute for Human-Machine Communication
Technische Universität München, Germany
cyril.joder@tum.de

Björn Schuller*
JOANNEUM RESEARCH DIGITAL - Institute for Information and Communication Technologies
Graz, Austria
bjoern.schuller@joanneum.at

ABSTRACT

This paper describes the system proposed for the Spoken Web Search task at Mediaeval 2012 campaign. We use an audio-only system based on our new called Cumulative Dynamic Time Warping (CDTW) algorithm. This algorithm combines the scores of all the alignment paths and allows for the learning of different cost functions for each kind of step in the alignment matrix (diagonal, horizontal and vertical). The results obtained with basic audio descriptors show the promising potential of our algorithm.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Spoken Term Detection

1. INTRODUCTION

The 'Spoken Web Search" (SWS) task of Mediaeval 2012 [2] consists in identifying specific words or expressions in spoken audio content, based on audio queries. The fact that the speech languages (four different South-African languages) are both unknown and resource-limited prevents the use of a dedicated speech recognition system. The goal of our submission is to assess the efficiency of our novel Cumulative Dynamic Time Warping (CDTW) algorithm.

1.1 Ressources

The present approach is audio-only and uses no external ressource.

2. CUMULATIVE DYNAMIC TIME WARPING (CDTW)

2.1 Definition

Cumulative Dynamic Time Warping (CDTW) is a novel variant of the well-known Dynamic Time Warping (DTW) algorithm for comparing two sequences. Several modifications are introduced compared to the standard algorithm.

First, the 'matching score' between the two sequences takes into account all the possible alignment paths instead of only the optimal one. This is done by replacing the 'hard' maximum (or minimum) operation of the DTW with the 'softmax' operation, defined as:

$$\text{softmax}(x_1, \ldots, x_n) = \log \sum_{i=1}^{n} e^{x_i}. \quad (1)$$

The other main change compared to common DTW is that the 'local distance' between frames of the two compared sequences is replaced with more general 'step scores', which depend on the local step taken. Formally, let $X = x_1, \ldots, x_I$ and $Y = y_1, \ldots, y_J$ be two vector sequences of lengths I and J respectively. The matching score $S(i,j)$ between the subsequences $X_{1:i} = x_1, \ldots, x_i$ and $Y_{1:j} = y_1, \ldots, y_j$ is calculated recursively as

$$S(i,j) = \text{softmax} \begin{cases} S(i-1,j-1) + s_1(i,j) \\ S(i-1,j) + s_2(i,j) \\ S(i,j-1) + s_3(i,j) \end{cases}, \quad (2)$$

wheres_1, s_2 and s_3 are the step scores associated to the 'diagonal', 'horizontal' and 'vertical' steps respectively. The overall normalised matching scores $\mathcal{S}(X,Y)$ between both sequences is then given by $\mathcal{S}(X,Y) = \frac{S(I,J)}{I+J}$.

2.2 Step Scores

The step scores are calculated as the weighted sum of several 'feature functions' $f_k(i,j)$, which characterize the local match between vectors x_i and y_j. Thus, the step scores can be written as: $s_m(i,j) = \sum_k \alpha_{m,k} f_k(i,j)$ for $m = 1, \ldots, 3$. Let D be the dimension of the considered vectors and let $x(d)$ be the d-th element of the vector x. In the present system, we use $5D+1$ feature functions, defined as follows:

$$\begin{aligned} \forall k = 1 \ldots D, \quad f_k(i,j) &= \big(x_i(k) - y_j(k)\big)^2 \\ f_{D+k}(i,j) &= \big(x_i(k) - x_{i-1}(k)\big)^2 \\ f_{2D+k}(i,j) &= \big(y_j(k) - y_{j-1}(k)\big)^2 \\ f_{3D+k}(i,j) &= \big(f_{D+k}(i,j) - f_{2D+k}(i,j)\big)^2 \\ f_{4D+k}(i,j) &= f_{D+k}(i,j) f_{2D+k}(i,j) \\ f_{5D+1}(i,j) &= 1. \end{aligned}$$

2.3 Classification

In order to classify if the two sequences are instances of the same phrase, the matching score is normalised and a logistic function is applied to obtain the decision function g:

$$g = \frac{1}{1 + e^{-(\beta \mathcal{S}(X,Y) - \delta)}}. \quad (3)$$

*This author is further affiliated with the Technische Universität München, Germany

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

Thanks to the softmax formulation of (2), the cumulative error function is derivable and a learning of the parameters $\alpha_{m,k}$, β and δ can be performed using gradient-based methods to minimize this cost function. For this system the learning was performed by stochastic gradient descent [1], with the constraint: $\forall k < D, \alpha_{m,k} \leq 0$. The training data was composed of part of the development data provided for the SWS task. Instances corresponding to the query data where extracted from the audio. Matching and Non-matching pairs where then exploited as positive and negative examples, respectively.

3. SEARCH PROCEDURE

The search for an occurrence of a query within the data is performed in three steps. First, the audio descriptors are extracted. Then, candidate matches are searched in each utterance. Finally, a matching score is calculated for each of the candidate matches.

In the present system the acoustic descriptors used are standard Mel-Frequency Cepstral Coefficients (MFCC) with log-energy, along with their first and second derivatives. The descriptors were extracted using HTK [3], with 25-ms windows and 10-ms hop-size. Cepstral mean normalization was then performed and each dimension was normalised to have zero mean and unitary variance.

For each audio utterance, the CDTW algorithm is run with an additional backtracking step, to compare the query with the whole utterance. The result is a matrix of matching scores corresponding to each pair of position i, j in the sequences. Next, a peak-picking process is performed on the first and last rows of this matrix, to extract potential starts j_{start} and ends j_{end} of matching sequences in Y. Then, match candidates are the selected sequences whose lengths are between $I/2$ and $2I$, where I is the length of the query.

Each match candidate is compared to the query thanks to the CDTW algorithm, resulting in a matching score $\mathcal{S}(X, Y_{j_{\text{start}}:j_{\text{end}}})$. We observed that using the same decision function as in (3) lead to a large number of false positive, as well as a bias towards some of the queries. Hence, a heuristic threshold $\delta(X)$, depending on the query, is defined as the 90-th percentile of the matching scores $\mathcal{S}(X, \cdots)$ in the candidate matches. The final decision function is then obtained by replacing the value δ with $\delta(X)$ in (3).

4. RESULTS

The evaluation metric for the task is the actual term weighted value (ATWV). Four runs have been performed, on the four combinations of query/audio collection data of development and evaluation sets (for example, evalQ-devC denotes the run with evaluation queries on development collection). The obtained AWTV, as well as probabilities of errors, are displayed in Table 1. Since we did not put much emphasis on tuning the threshold, we also consider the maximum ATWV (obtained with optimal threshold Thr), represented in Figure 1 with the whole Detection Error Tradeoff (DET). These results confirm the potential of CDTW compared to standard DTW. Indeed, replacing DTW with CDTW for comparing the candidate matches in the described system yielded a maximum ATWV of 0.065 for the dev/dev run (instead of 0.270). We can also observe that, although the actual values of the maximum ATWV vary between the runs, the DET curves are quite similar, indicating that our system generalizes well to unknown data. However, the value of the optimal threshold also varies. This shows that a more elaborate decision-making step is needed for a better system.

run	P(miss)	P(FA)	ATWV
devQ-devC	55.6 %	1.18 %	0.263
evalQ-devC	59.5 %	1.13 %	0.333
devQ-evalC	60.2 %	1.17 %	0.164
evalQ-evalC	54.5 %	1.13 %	0.290

Table 1: Results obtained in the four runs. P(miss) and P(FA) denote the probabilities of false negative and false positive respectively.

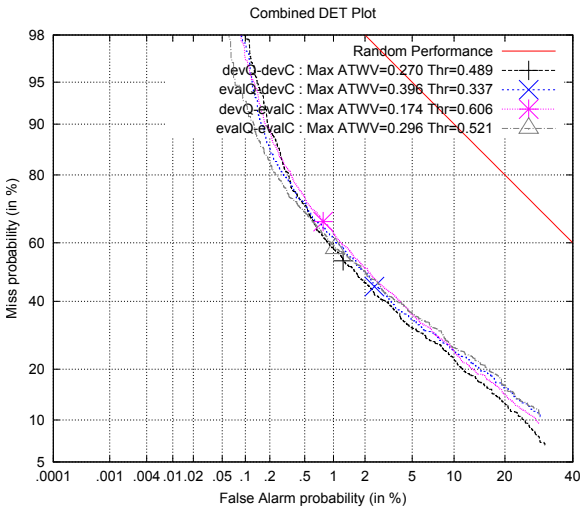

Figure 1: DET curves and maximum ATWV.

5. CONCLUSION

The main purpose of our submission was to assess the usefulness of the CDTW algorithm to compare sequences. The obtained results are promising, since our system shows a significant improvement over standard DTW as well as good generalization property. However, many improvement can still be imagined, such as the design of a more elaborate decision function. Moreover, although the appending of English phoneme posterior probabilities has been attempted without showing significant improvements, more robust acoustic descriptors should allow for further developments.

6. REFERENCES

[1] L. Bottou. Stochastic learning. In O. Bousquet and U. von Luxburg, editors, *Advanced Lectures on Machine Learning*, Lecture Notes in Artificial Intelligence, LNAI 3176, pages 146–168. Springer Verlag, Berlin, 2004.

[2] F. Metze, E. Barnard, M. Davel, C. van Heerden, X. Anguera, G. Gravier, and N. Rajput. The Spoken Web Search Task. In *MediaEval 2012 Workshop*, Pisa, Italy, October 4-5 2012.

[3] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK, 2006.