TUKE MediaEval 2012: Spoken Web Search using DTW and Unsupervised SVM

Jozef Vavrek
KEMT, FEI, Technical
University of Košice
Park Komenského 13
041 20 Košice, Slovakia
Jozef.Vavrek@tuke.sk

Matúš Pleva
KEMT, FEI, Technical
University of Košice
Park Komenského 13
041 20 Košice, Slovakia
Matus.Pleva@tuke.sk

Jozef Juhár
KEMT, FEI, Technical
University of Košice
Park Komenského 13
041 20 Košice, Slovakia
Jozef.Juhar@tuke.sk

ABSTRACT

This working paper provides the basic information about experiments conducted on audio documents within the MediaEval 2012 spoken web search evaluation project. The main purpose of these experiments was to build a robust and language independent system for spoken term detection. Therefore we have proposed query-by-example searching system based on the minimum-cost alignment of DTW algorithm and unsupervised SVM misclassification rate. Results show that our system is liable to variable length of queries with similar spectral characteristics that results in poor detection performance with high number of insertions and misdetections. There were no other resources used during the development.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Spoken Term Detection - term matching algorithm based on unsupervised SVM and DTW

Keywords

Query-by-example search, Support Vector Machines, Dynamic Time Warping

1. MOTIVATION

The Spoken Web Search task of MediaEval 2012 [3] involves searching for audio content within audio content database using an audio content query. This very challenging task is very useful for applying and testing new audio content searching and classification algorithms and their combinations which were also used in our previous work on broadcast news classification task[4]. Our research also covers an acoustic events detection task, which involves a support system for CCTV camera operators, where different feature extraction algorithms were tested and evaluated for searching acoustic query in acoustic online stream [5].

2. SYSTEM OVERVIEW

The basic computational architecture of our proposed Spoken Term Detection (STD) system consists of two main parts. The first part deals with features that objectively represents audio-content of queries and utterances. The main functionality of the second part is to implement techniques that help to locate occurrences of each query within the utterances.

2.1 Parameterization technique

The substantial step in comparing audio content of two segments is to extract parameters (coefficients) that capture temporal and spectral characteristics of the audio signal. The process of feature extraction has a major impact for further pattern matching and similarity measure techniques. It helps to reduce the probability of misdetection and improves the retrieval performance within the audio document so that system is able to detect each specific query over all utterances with minimum false detection rate. Therefore we have experimented with several parameterization techniques, such as MFCCs, ZCR and MPEG-7 low level descriptors (ASS, ASC, ASF, ASE). The main effort was aimed at selecting only a particular set of features that are able to capture required characteristics of queries, in order to match corresponding terms within the utterances. We have tried several combinations of the features mentioned above and chose the one with the average minimum-cost alignment (avgMCA), by using DTW algorithm, between selected queries and corresponding terms within utterances.

The optimal results were achieved by using first 12 MFCCs features and zero energy coefficient (the avgMCA was about 250.1). It follows that the value of avgMCA increases rapidly with the number of feature dimension which causes higher computational cost in the process of training the SVM classifier. Therefore we decided to use only MFCCs in our searching algorithm with the window length of $l_{window} = l_{query}/100$, where l_{query} is the length of each individual query. The frame-shift was set as 50% of window length. The main reason we have used variable window length for each frame within the all utterances was the big differences in durations of each particular query. It means that there would be a large number of frames in case of queries with long duration and only few frames for queries with very short duration. The next reason was the minimal number of instances the SVM classifier needs in order to be effective during the training phase. The big disadvantage of using a variable window length is the relatively high computational cost, because we need to compute MFCCs for each particular query and utterances separately.

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

2.2 Query-by-example search algorithm using DTW and Unsupervised SVM

The main functionality of our proposed search algorithm lies in comparing two audio segments, with the same length, by using DTW algorithm and misclassification rate of Support Vector Machine (SVM) classifier. The first segment represents the searching query and the second one refers to the audio segment of each utterance with the same length as the query segment [2]. The system compares only two segments at one time. The first segment is the same for the same query and second one is shifted to the right until the end of the utterances audio file is reached. Subsequently, the minimum-cost alignment (MCA) between each couple of segments was computed by using DTW algorithm. Then we chose four segments, within the one utterances audio file, with minimal value of MCA and assumption that the lower value of MCA is the higher probability of detection the corresponding segment. This process was then repeated for each search query within every provided audio file.

The same process of shifting segments according to the length of te search query was then applied for computing the misclassification rate of SVM classifier. Our idea was based on the assumption that if two segments, which represent the same class (+1) or (-1), contain the same spoken word then they do not have significant differences and an SVM hyperplane would not be able to effectively classify these segments into two classes. On the contrary, two segments with different spoken word represent different classes (+1) and (-1) have significant differences so that the SVM hyperplane can effectively classify these two classes. It follows that the number of iteration in the process of finding the optimal decision hyperplane increases with the same classes and decreases with two different classes. Therefore, the substance of unsupervised learning lies in defining only two classes in the process of training the SVM classifier. So that our system does not need to define number of all possible acoustic classes (several words spoken by different speakers). This makes this system language independent.

Fei et al. [1] have used similar procedure for detecting speaker changes, based on a new SVM training misclassification rate. They have carried out several experiments in order to set the best parameter for training the SVM. So that we decided to use *linear* kernel function, penalty parameter $C = 1$ and segment shift $l_{query}/2$. The latter parameter was set by us after several experiments. Misclassification rate is defined as the number of frames misclassified from one class to another. It is computed from predicted data.

3. EVALUATION AND CONCLUSIONS

All experiments were carried out on African data set [2]. The overall procedure of searching algorithm is depicted in Fig. 1. Each step of the algorithm is repeated until the last query is reached within the utterance files. The misclassification rate and predicted accuracy of the SVM were tuned experimentally after several runs of the algorithm. Misclassification rate for two classes (miss(+1) and miss(-1)) was computed for predicted data and threshold set to 0.12. Predicted accuracy is defined as number of correctly predicted instances to all tested instances. The searching query was observed if predicted data error rate (1-accuracy) within all segments and utterances was higher than 18%.

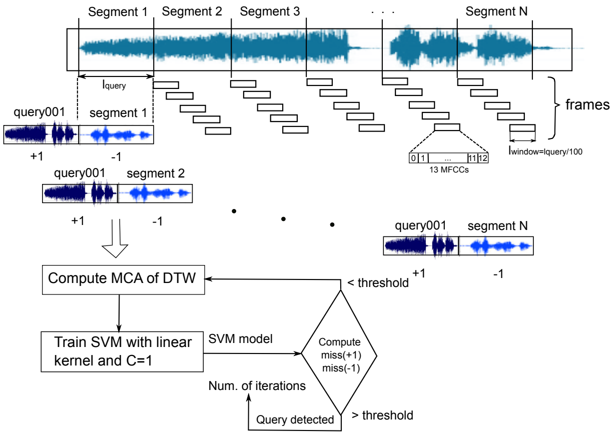

Figure 1: Basic architecture of searching algorithm using DTW and unsupervised SVM

Table 1: Evaluation results of the tested algorithm

database set	*P(FA)*	*P(Miss)*	*ATWV*
evalQ-devC	1.54617	0.960	-0.052
devQ-evalC	1.62595	0.948	-0.233
evalQ-evalC	1.68694	0.974	-0.164
devQ-devC	1.78786	0.943	-0.194

We have found out that our algorithm produced a huge number of spurious term detections during evaluation. The overall performance of our system was evaluated on different sets by probability of false alarms P(FA), misdetection P(Miss) and the actual term weighted value ATWV. The results are listed in Tab. 1. It follows that the algorithm detected correctly only 6% terms with $22k$ false alarms and -0.194 ATWV, in case of devQ-devC, while the score threshold was set on 2.82 (25% with $45k$ FA using 2.86).

4. ACKNOWLEDGMENTS

The research was supported by the Ministry of Education of the Slovak Republic under project VEGA 1/0386/12 (50%) & ERDF funded ITMS-26220220155 project (50%).

5. REFERENCES

[1] B. Fei and J. Liu. Unsupervised speaker change detection using SVM training misclassification rate. *IEEE Trans. Neural Netw.*, 17:696–704, May 2006.

[2] A. Kumar, N. Rajput, D. Chakraborty, S. K. Agarwal, and A. A. Nanavati. WWTW: the world wide telecom web. In *NSDR 2007 (SIGCOMM workshop)*, pages 84–89, Kyoto, Japan, August 27 2007.

[3] 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.

[4] J. Vavrek, E. Vozáriková, M. Pleva, and J. Juhár. Broadcast news audio classification using SVM binary trees. In *TSP Proceedings*, pages 469–473, July 2012.

[5] E. Vozáriková, M. Pleva, S. Ondáš, J. Vavrek, J. Juhár, and A. Čižmár. Detection and classification of audio events in noisy environment. *Journal of Computer Science and Control Systems*, 3(1):253–258, 2010.