The NNI Query-by-Example System for MediaEval 2014

Peng Yang[1], Haihua Xu[2], Xiong Xiao[2], Lei Xie[1], Cheung-Chi Leung[3], Hongjie Chen[1], Jia Yu[1], Hang Lv[1], Lei Wang[3], Su Jun Leow[2], Bin Ma[3], Eng Siong Chng[2], Haizhou Li[2,3]
[1]Northwestern Polytechnical University, Xi'an, China
[2]Nanyang Technological University, Singapore
[3]Institute for Infocomm Research, A*STAR, Singapore
pengyang@nwpu-aslp.org, haihuaxu@ntu.edu.sg, xiaoxiong@ntu.edu.sg

ABSTRACT

In this paper we describe the system proposed by NNI (NWPU-NTU-I2R) team for the QUESST task within the Mediaeval 2014 evaluation. To solve the problem, we used both dynamic time warping (DTW) and symbolic search (SS) based approaches. The DTW system performs template matching using subsequence DTW algorithm and posterior representations. The symbolic search is performed on phone sequences generated by phone recognizers. For both symbolic and DTW search, partial sequence matching is performed to reduce missing rate, especially for query type 2 and 3. After fusing 9 DTW systems, 7 symbolic systems, and query length side information, we obtained 0.6023 actual normalized cross entropy (actCnxe) for all queries combined. For type 3 complex queries, we achieved 0.7252 actCnxe.

1. INTRODUCTION

This paper presents the NNI team's system for the Query-by-Example Search on Speech task (QUESST) within the Mediaeval 2014 evaluation [1]. Our system is a fusion of 2 groups of component systems as shown in Fig. 1 (diagram inspired by [2]). One group is based on sub-sequence DTW which worked well on exact match task (query type 1) in the previous Spoken Web Search tasks [3, 4]. The other group is based on symbolic search, i.e. to match phone sequences of queries and search data. The symbolic search used in this paper was motivated by the Open Keyword Search task [5].

2. TOKENIZERS

We used several phone recognizers trained from different resources as tokenizers, including 3 BUT phone recognizers [6] (Czech, Hungarian, and Russian), 3 phone recognizers trained from the Switchboard corpus [7] (1 triphone DNN model, 1 monophone DNN model, and 1 stacked bottleneck feature (SBN) based GMM model), and 2 phone recognizers trained from a Malay corpus [8] (triphone and monophone DNN models). These tokenizers were used for both DTW and symbolic systems. The SBN features were used in a DTW system directly. Besides phone recognizers, we also trained a 1024-component Gaussian mixture model (GMM) from the search data using VTLN-processed MFCC features as in [9]. The Gaussian posteriors are used as the input of a DTW system.

3. DTW-BASED APPROACH

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

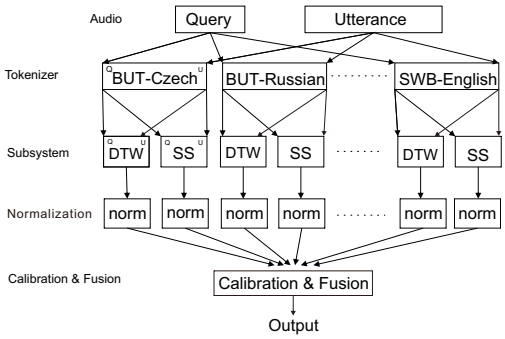

Figure 1: Diagram of the NNI QbE-STD system.

We implemented 9 DTW systems - 5 full matching and 4 partial matching systems. Full matching means that a system computes a score by aligning the complete sequence of query features (after VAD) with a test utterance. The average accumulated distance of the best aligned path between 2 sentences is obtained in a dynamic programming manner [4]. The 5 full matching systems used phoneme state posteriorgram (BUT Russian, Czech, and Hungarian), Gaussian posteriorgram and SBN as features. Inner-product distance was used for posteriorgram features and cosine distance was used for non-posteriorgram features.

To deal with type 3 queries, we also implemented 4 partial matching systems (CZ, HU, RU, and SBN), where only part of the query features is used for matching. The aim is to detect instances of queries based on only partial evidence in the sentences. This approach share the same spirit as the partial symbolic sequence matching to be discussed in next section and was also inspired by [10]. In implementation, we obtained a set of feature segments for each query by shifting a 600ms window every 50ms along the query features. All feature segments are matched with a test utterance and the best score is chosen to represent the query-sentence trial. All DTW scores were normalized to have zero mean and unit variance per query.

4. SYMBOLIC SEARCH

The symbolic search used here is motivated by the weighted finite state transducer (WFST) based keyword search (KWS) system that is popular in the OpenKWS tasks [5]. However, there are quite a lot of differences between the QUESST task and the OpenKWS task. In OpenKWS, only one language is searched and labeled training data is provided to train an LVCSR system. Hence search data is recognized into word/phone lattice which is converted to WFST format for easy search. Text queries are converted to WFST

also. A composition of query and search data WFSTs will return the common paths of the 2 WFSTs, and hence the exact match.

To apply the WFST based framework to the QUESST task, 2 modifications were made. First, as there is no labeled data to train a LVCSR system, we must rely on the phone recognizers trained from other resources. In this task, we used 2 BUT phone recognizers (Czech and Hungarian), and all the 3 Switchboard and 2 Malay phone recognizers described in section 2. Second, the audio query is also converted into phone sequences. The matching of query to search data is performed by composing query and data WFSTs.

Due to the high variation in the queries and search data, e.g. different recording channels and languages, the symbolic search that uses exact match performs poorly on the QUESST data. The missing rate is very high as the phone representations of the query and search data can be very different. To reduce missing rate, we used top-N hypotheses for each query, where N can be as high as 1000. All top-N hypotheses were treated equally and searched. Hypotheses shorter than 5 phones were discarded to avoid false alarms from exploding. The number of hypotheses N depends on query length. If a query has M phones, 2^M hypotheses will be used as more hypotheses are needed to adequately represent the variations of longer queries. Besides using top-N hypotheses, we also used partial phone sequence matching to further reduce the missing rate of long queries. Long queries (e.g. longer than 8 phones) are difficult to be detected, even using up to 1000 hypotheses, due to the exponentially increasing of hypotheses variations with query length. To address this problem, we used all partial phone sequences of the queries in search. For example, we found that using all partial sequence of length 6 worked well on the QUESST dev data. The partial sequence matching also make it easier to detect type 2 and type 3 queries. For example, in type 3 queries, we are required to return "horse white" if the query is "white horse". If partial sequence matching used, the system will return a hit once it detects "horse" or "white". Although the partial matching has the potential of increasing false alarm, we found that for the QUESST data, it worked well on type 3 queries.

5. EXPERIMENTAL RESULTS

The results of the proposed system on the QUESST evaluation are listed in Table 1. In the table, "DTW" refers to the fusion of 9 DTW systems using Focal [11]. "Symbolic" is the fusion of 7 SS based systems. "Fusion" refers to the fusion of all 9 DTW and 7 SS systems. "Fusion+Length" means adding query length information to the fusion. From the results, the overall performance of symbolic search is slightly worse than the DTW system. When we look into individual query types, DTW has a big advantage over symbolic search in type 1 queries. For type 2 queries, the two approaches perform similarly. For type 3, symbolic search has big advantage on TWV, but not Cnxe. The observation may be attributed to the fact that the symbolic search systems rely heavily on partial matching, which is suitable for type 3 queries. Another observation is that adding query length produces better Cnxe, but worse TWV. This could be due to that the fusion is optimized on Cnxe, not TWV.

The peak memory usage of all DTW systems is 60GB when all feature representations are loaded, and the searching speed factor (SSF) is 0.1054. Each SS system takes 80 CPU hours to index the 23 hours of search audio (ISF=3.5) and 50 hours to search 555 eval queries (single system SSF=0.0012, 7-system SSF=0.0085). Peak memory during search is 45GB per SS system.

6. CONCLUSIONS

We have described the NNI system for the QUESST 2014 task.

Table 1: Performance of DTW and Symbolic search on eval data. Results are separated by query types.

Methods	Cnxe	MinCnxe	ATWV	MTWV
Type 1 Queries				
DTW	0.5733	0.5971	0.4448	0.4465
Symbolic	0.6787	0.6715	0.3526	0.3603
Fusion	0.5248	0.5088	**0.5115**	**0.5136**
Fusion+Length	**0.5074**	**0.4946**	0.4989	0.5010
Type 2 Queries				
DTW	0.7300	0.7191	0.2306	0.2408
Symbolic	0.7405	0.7338	0.2294	0.2357
Fusion	0.6386	0.6290	0.3158	**0.3324**
Fusion+Length	**0.6212**	**0.6144**	**0.3205**	0.3234
Type 3 Queries				
DTW	0.8029	0.7925	0.1465	0.1673
Symbolic	0.8035	0.7950	0.2134	0.2237
Fusion	**0.7210**	0.7140	**0.3061**	**0.3102**
Fusion+Length	0.7252	**0.7100**	0.2871	0.2925
All Queries				
DTW	0.6925	0.6816	0.2918	0.2974
Symbolic	0.7322	0.7293	0.2696	0.2717
Fusion	0.6125	0.6062	**0.3896**	**0.3952**
Fusion+Length	**0.6023**	**0.5977**	0.3792	0.3801

We have leveraged on both the advantage of DTW systems on type 1 queries, and partial matching symbolic search systems on type 3 queries. The partial patching strategy used in both symbolic and DTW systems helps to reduce missing rate significantly, especially for type 3 queries. Future research will be focused on reducing the false alarms introduced by partial matching.

7. REFERENCES

[1] Anguera X., Rodriguez-Fuentes L. J., Szöke I., Buzo A., and Metze F., "Query by example search on speech at mediaeval 2014," in *Working Notes Proceedings of the Mediaeval 2014 Workshop*, Barcelona, Spain, Oct. 16-17.

[2] Szoke I. et al., "Calibration and fusion of query by example systems-BUT SWS 2013," in *Proc. ICASSP*, 2014.

[3] Rodriguez-Fuentes L. J. et al., "High-performance query-by-example spoken term detection on the sws 2013 evaluation," in *Proc. ICASSP*, 2014.

[4] Yang P. et al., "Intrinsic spectral analysis based on temporal context features for query-by-example spoken term detection," in *Proc. Interspeech*, 2014.

[5] OpenKWS13, "Openkws13 keyword search evaluation plan," *available online: http://www.nist.gov/itl/iad/mig/upload/OpenKWS13-EvalPlan.pdf.*

[6] Schwarz P. et al., "Hierarchical structures of neural networks for phoneme," in *Proc. ICASSP*, 2006.

[7] Godfrey J. J. et al., "Switchboard: Telephone speech corpus for research and development," in *Proc. ICASSP*, 1992.

[8] Tan T. P. et al., "Mass: A Malay language LVCSR corpus resource," in *Proc. O-COCOSDA*, 2009.

[9] Wang H. et al., "An acoustic segment modeling approach to query-by-example spoken term detection," in *Proc. ICASSP*, 2012.

[10] Zheng L. et al., "Acoustic texttiling for story segmentation of spoken documents," in *Proc. ICASSP*, 2012.

[11] Brummer N., "Focal toolkit," in *https://sites.google.com/site/nikobrummer/focal.*