IIIT-H System for MediaEval 2014 QUESST

Santosh Kesiraju, Gautam Mantena, Kishore Prahallad
International Institute of Information Technology-Hyderabad, India
{santosh.k, gautam.mantena}@research.iiit.ac.in, kishore@iiit.ac.in

ABSTRACT

This paper describes the experiments and observations for Query-by-Example Search on Speech Task (QUESST) at MediaEval 2014. In this paper, we describe two different representations of speech that were explored for the task. We also show the capabilities and limitations of non-segmental dynamic time warping (NS-DTW) technique for searching various types of queries. This paper mainly focuses on the experiments and analysis of the existing NS-DTW algorithm for various types of queries. The observations show that for a specific representation of speech, the algorithm is capable of detecting partial matches.

1. INTRODUCTION

Some of the approaches for query-by-example spoken term detection rely on building models from resource rich languages, and use these models to convert the speech data into sequence of symbols. Building models for multi-lingual data is a challenging task as phone classes are not language universal. Another way is relying on dynamic time warping (DTW) based techniques for matching two time series vectors. Here, speech data is usually represented as Gaussian posteriorgrams (GP) of various acoustic features.

For MediaEval 2014 QUESST task [2], we have explored unsupervised techniques involving various representations for the speech data. Initially, we represented the speech data using GP of acoustic and bottle-neck features. We have also built a cross-lingual ASR and decoded the speech data into a sequence of symbols (phone sequences). Both the representations rely on DTW to detect the queries in the audio references.

2. FEATURE REPRESENTATION

A three step process to generate the features for queries and the audio references is described here. (a) 39 dimensional frequency domain linear prediction (FDLP) features along with delta and acceleration coefficients were extracted for every 25 ms window and a shift of 10 ms. An all-pole model of order 160 poles/sec and 37 filter banks were considered to extract FDLP features. (b) Bottle neck (BN) features were derived from Multi-layer perceptron (MLP) trained with articulatory features (AF) (c) Gaussian posteriorgrams were computed for speech parameters (FDLP) in tandem with articulatory bottle neck features. Bottle neck features are a form of compressed features which are of lower dimension and also capture the classification properties of the target classes. These features were obtained from the MLP trained on 24 hours of labeled Telugu database [3]. The articulatory bottle neck features were extracted as described in [5].

3. NS-DTW FOR SEARCH

We used a variant of DTW called non-segmental DTW (NS-DTW) [4], which differs in the local constraints. As a post processing method, we have pruned out some of the results. The pruning criteria is based on the slope of the aligned path. If m is the slope of the aligned path, then, only the paths satisfying ($0.5 < m < 2$), were considered. This helped us in eliminating some of the false alarms. We have used the linear calibration function in bosaris toolkit [1] to calibrate the scores. Table 1 shows the results on development and evaluation dataset for different types of queries.

Table 1: Scores for various types of queries for (FDLP + AF-BN) feature representation on dev and eval datasets

dev dataset				
	Type of queries			
Scores	All	Type 1	Type 2	Type 3
MinCnxe	0.8070	0.6734	0.8739	0.8986
Cnxe	0.9121	0.8032	1.0121	1.0235
MTWV	0.2263	0.3715	0.1472	0.0430
ATWV	0.2261	0.3662	0.1467	0.0425
eval dataset				
	Type of queries			
Scores	All	Type 1	Type 2	Type 3
MinCnxe	0.8117	0.7006	0.8576	0.8936
Cnxe	0.9218	0.8115	1.0205	1.0012
MTWV	0.2062	0.3506	0.1188	0.0770
ATWV	0.2026	0.3475	0.1151	0.0655

All the experiments were performed on a single HP SL230 node which is equipped with two Intel E5-2640 processors with 12 cores each and 64 GB of main memory. The peak memory usage (PMU) was approximately 12 GB. The searching speed factor (SSF) was 3.46.

To increase the search speed, the distance computation was parallelized on a GPU (NVIDIA GT 610 with 48 cores and 2 GB of GPU memory). The SSF was reduced to 0.85.

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

[1] https://sites.google.com/site/bosaristoolkit/

4. ANALYSIS OF THE EXPERIMENTS

We have analyzed the cases of false alarms and misses for all types of queries. The analysis on false alarms helped us in enforcing a slope constraint on the aligned path which was described in Section 3. The results in Table 1 show that the NS-DTW algorithm is able to detect some of the type 2 queries, but fails in detecting type 3 queries. Fig. 1(a) shows the similarity matrix plot for a multi-word query with filler content present in the reference. The dark bands represent the match between the query and the reference. In Fig. 1(a) there are multiple dark bands, each showing a match between parts of the query (word) to the specific locations (words) in the reference. The peaks in the alignment scores in Fig. 1(b) reflects the partial matches. This shows that for this specific (FDLP + AF-BN) feature representation of speech, the algorithm is capable of detecting smaller/partial matches. Even though the scores reflect the partial matches, we have observed that the poor performance of the system is due to the number of false alarms. Further investigation is required to find the methods that can penalize the false alarms.

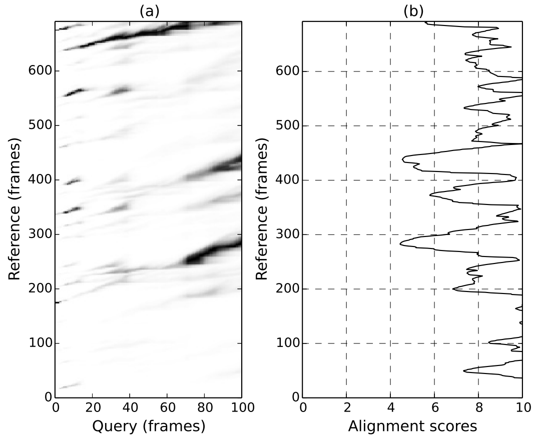

Figure 1: An example similarity matrix obtained using NS-DTW, when multi-word query with filler content is present in the reference.

5. USING PHONE DECODER

In this work, we have also built a cross-lingual phone decoder and used NS-DTW for search. The cross-lingual decoder was built in a two step process. As the first step, we trained acoustic models on 24 hours of Telugu database [3]. Then these models were used to decode MediaEval 2013 SWS database [1]. The decoded symbols were bootstrapped and the models were re-trained. This process was repeated 4 times and the resulting acoustic models were used to obtain the hypotheses (global hypotheses).

We have built a phone confusion matrix in an unsupervised way which is as follows: (a) We divided the SWS 2013 database into 4 parts and 4 acoustic models were built (b) 4 hypotheses (local hypotheses), each corresponding to a different part of the database were obtained (c) A string alignment was done between the global hypotheses and each of the local hypotheses to obtain the phone confusions. The global hypotheses was considered as the reference in computing the phone confusions. Next, the queries and the audio references were decoded using the bootstrapped models, and the search was performed using the NS-DTW. The phone confusion matrix was used in the computation of similarity matrix in the NS-DTW framework.

If i and j are the indices of phones and N is the number of phones in the dictionary, then the similarity between them is given by,

$$d(i, j) = c(i, j) \quad \forall \quad 0 \leq i, j \leq N$$

where $c(i, j)$ is the confusion matrix of i being the reference phone and j being the query phone.

The SSF in this case was 0.38 and the PMU was approximately 2 GB. The results for various types of queries on development dataset are shown in Table 2.

Table 2: Scores for various types of queries for phone representation on dev dataset

Phone representation				
	Type of queries			
Scores	All	Type 1	Type 2	Type 3
MinCnxe	0.9487	0.9331	0.9599	0.9641
MTWV	0.0477	0.0799	0.0308	0.0134

6. CONCLUSION

In this work, we have explored two different representations of speech. We have observed the capabilities and limitations of NS-DTW algorithm for various types of queries. We have also observed that the same algorithm is able to detect some of the type 2 queries in the reference documents. The future work is focused on improving the NS-DTW algorithm for detecting type 2 and type 3 queries and also in developing robust cross-lingual phone decoders.

7. REFERENCES

[1] X. Anguera, F. Metze, A. Buzo, I. Szoke, and L. J. Rodriguez-Fuentes. The Spoken Web Search Task. In *Working Notes Proceedings of the MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[2] X. Anguera, L. J. Rodriguez-Fuentes, I. Szoke, A. Buzo, and F. Metze. Query by Example Search on Speech at Mediaeval 2014. In *Working Notes Proceedings of the Mediaeval 2014 Workshop*, Barcelona, Spain, October 16-17 2014.

[3] G. K. Anumanchipalli, R. Chitturi, S. Joshi, S. S. R. Kumar, R. Sitaram, and S. Kishore. Development of Indian language speech databases for LVCSR. In *Proc. of SPECOM*, Patras, Greece, 2005.

[4] G. Mantena, S. Achanta, and K. Prahallad. Query-by-example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(5):946–955, May 2014.

[5] G. Mantena and K. Prahallad. Use of articulatory bottle-neck features for query-by-example spoken term detection in low resource scenarios. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7128–7132, May 2014.