CUHK System for QUESST Task of MediaEval 2014

Haipeng Wang, Tan Lee
DSP-STL, Dept. of EE
The Chinese University of Hong Kong
{hpwang,tanlee}@ee.cuhk.edu.hk

ABSTRACT

This paper describes a spoken keyword search system developed at the Chinese University of Hong Kong (CUHK) for the query by example search on speech (QUESST) task of MediaEval 2014. This system utilizes posterior features and dynamic time warping (DTW) for keyword matching. Multiple types of posterior features are generated with different tokenizers, and then fused by a linear combination on the DTW distance matrices. The main contribution of this year's system is a multiview segment clustering (MSC) approach for unsupervised ASM tokenizer construction. The C_{nxe} and ATWV of our submitted results on the Evaluation set are 0.682 and 0.412, respectively.

1. INTRODUCTION

The query by example search on speech (QUESST) task aims at detecting the keyword occurrences in a unlabeled speech collection using spoken queries in a language independent fashion. In this year's QUESST dataset, the speech collection involves about 23 hours of speech data from 6 languages, and the query set includes 560 development queries and 555 evaluation queries. The average duration of queries is about 0.9 second after voice activity detection (VAD). More details about the task description can be found in [2].

Our system was designed only for the type 1 query matching. It followed the posteriorgram-based template matching framework [3], in which speech tokenizers were used to generate posteriorgrams, and DTW was applied for keyword detection. The tokenizers were either built from the searching speech collection given in the task, or developed from some resource-rich languages. In order to exploit the complementary information of multiple tokenizers, the DTW matrix combination method [7] was used. Raw DTW detection scores were then normalized to zero mean and unit variance. On the evaluation set, the C_{nxe} and ATWV of our submission are 0.682 and 0.412. If only considering the type 1 query matching, the C_{nxe} and ATWV are 0.611 and 0.526.

2. SYSTEM DESCRIPTION

2.1 System Overview

In this year's evaluation, our system employs a similar framework as our previous system for spoken web search task in 2012 [5]. The system involves seven tokenizers, including a GMM tokenizer, five phoneme recognizers, and an ASM tokenizer [8]. Using these tokenizers, the query examples and test utterances are converted into frame-level posteriorgrams. Different tokenizers may use different algorithms to generate posteriorgrams. Let $\mathbf{Q}_i$ denote the query posteriorgram generated by the i_{th} tokenizer, and let $\mathbf{T}_i$ denote the corresponding test posteriorgram. The distance matrix $\mathbf{D}_i$ was computed as the inner-product [3],

$$\mathbf{D}_i = -\log(\mathbf{Q}_i^T \times \mathbf{T}_i) \quad i = 1, 2, ..., 7. \tag{1}$$

To exploit the complementary information from different tokenizers, the distance matrices were combined linearly to give a new distance matrix $\mathbf{D}$,

$$\mathbf{D} = \sum_{i=1}^{7} w_i \mathbf{D}_i, \tag{2}$$

where w_i denotes the weighting coefficients for $\mathbf{D}_i$ and was simply set to $\frac{1}{7}$.

Subsequently, DTW detection was applied to the combined distance matrix $\mathbf{D}$ to locate the top matching regions. DTW detection was performed with a sliding window with a window shift of 5 frames. The adjustment window constraint was imposed on the DTW alignment path. Let $d_{q,t}$ denote the normalized DTW alignment distance between the q_{th} query on the t_{th} hit region. The raw detection score was computed by

$$s_{q,t} = \exp(-d_{q,t}/\beta), \tag{3}$$

where the scaling factor β was set to 0.6. To calibrate the score distribution of different queries, a 0/1 normalization was used,

$$\hat{s}_{q,t} = (s_{q,t} - \mu_q)/\delta_q, \tag{4}$$

where $\hat{s}_{q,t}$ is the calibrated score, and μ_q and δ_q^2 are the mean and variance of the raw scores of the q_{th} query.

2.2 GMM Tokenizer

The GMM tokenizer was trained from the given searching speech collection. It contained 1024 Gaussian components. The input of the GMM tokenizer was 39-dimensional MFCC feature vector. The MFCC features were processed with VAD and utterance-based mean and variance normalization (MVN). Vocal tract length normalization (VTLN) was then applied to the MFCC features o alleviate the influence of speaker variation.

The warping factors of VTLN were estimated iteratively as proposed in [9]. The iteration started with training a

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

GMM from the unwarped MFCC features. Then the warping factors were estimated with a maximum-likelihood grid search using the GMM. A new GMM was trained using the warped features, and new warping factors were then re-estimated. This process was iterated four times in our implementation. The usefulness of VTLN for this task was experimentally demonstrated in our previous paper [8].

2.3 Phoneme Recognizers

Our system involved five phoneme recognizers, namely Czech, Hungarian, Russian, English and Mandarin phoneme recognizers. All these phoneme recognizers used the split temporal context network structure [4]. The Czech, Hungarian, Russian phoneme recognizers were developed at Brno University of Technology (BUT) and released in [1]. The English phoneme recognizer was trained on about 15-hour speech data from the Fisher corpus and Swichboard Cellular corpus. The Mandarin phoneme recognizer was trained on about 15-hour speech data from the CallHome corpus and the CallFriend corpus. These phoneme recognizers were used to generate mono-phone state-level posteriorgrams without any language model constraint.

2.4 ASM Tokenizer

Acoustic segment modeling (ASM) is a way to build an HMM-based speech tokenizer from unlabeled speech data. It consists of three steps, namely initial segmentation, segment labeling, and iterative training and decoding. Initial segmentation searches for the acoustic discontinuities and partitions speech utterances into short-time speech segments. In our implementation, we simply used the one-best recognition results of the Hungarian phoneme recognizer to get the hypothesised segment boundaries.

Segment labeling is to assign a label to each short-time speech segment. We used a multiview segment clustering (MSC) approach for segment labeling. The MSC approach took in multiple segment-level posterior features, computed the similarity matrix and Laplacian matrix of the speech segments for each type of posterior feature, and made a linear combination on the Laplacian matrices. With the combined Laplacian matrix, eigen-decomposition was performed to derive the spectral embedding representations, and k-means was applied to find 100 clusters. Details of the MSC approach are described in [6].

The cluster labels were used as initializations for iterative training and decoding, in which HMM training and decoding were performed iteratively until converge.

3. RESULTS

Table 1 shows the results obtained by our system on evaluation queries. Based on our previous experience on TWV values, we only submitted a small portion of the scores which were higher than a threshold. This gives us the results of System 1. However, if all the scores of all the trials are considered, we obtain the results of System 2, which gives obvious reductions on the C_{nxe} values. Similar observations can also be made when only considering the type 1 query matching. Corresponding results are shown in Table 2. The difference between C_{nxe} and TWV metrics needs to be carefully examined in the future.

To run the experiments, we used a computer with Intel i7-3770K CPU (3.50GHz, 4 cores), 32GB RAM and 1T hard drive. In the online searching process, all the posteriorgrams were stored in the memory. This caused very high memory cost (>10GB). The computation cost in the searching process was mainly caused by DTW detection. The searching speed factor of our system was about 0.021. The slow searching speed is one main drawback of our system and needs to be improved.

Table 1: System performances on all the queries. System 1 corresponds to the submitted results.

System No.	actC_{nxe}	minC_{nxe}	ATWV	MTWV
1	0.682	0.659	0.412	0.413
2	0.638	0.585	0.412	0.413

Table 2: System performances on the type 1 queries. System 1 corresponds to the submitted results.

System No.	actC_{nxe}	minC_{nxe}	ATWV	MTWV
1	0.526	0.486	0.611	0.613
2	0.508	0.420	0.611	0.613

4. CONCLUSION

We have described an overview of the CUHK system submitted to the MediaEval 2014 QUESST task along with the evaluation results. Our system involves seven tokenizers and uses DTW matrix combination for fusion. Only type 1 query matching is considered in the system development. The main highlight of our system lies in the MSC approach in the ASM tokenizer construction. In general we think the performances for type 1 query matching are acceptable, but the slow searching speed and high memory cost need to be substantially improved.

5. REFERENCES

[1] http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context.

[2] X. Anguera, L. Rodriguez-Fuentes, A. B. I. Szoke, 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] T. Hazen, W. Shen, and C. White. Query-by-example spoken term detection using phonetic posteriorgram templates. In *ASRU*, pages 421–426, 2009.

[4] P. Schwarz. Phoneme recognition based on long temporal context, PhD thesis. 2009.

[5] H. Wang and T. Lee. CUHK system for the spoken web search task at mediaeval 2012. In *Working Notes Proceedings of the Mediaeval 2012 Workshop, Pisa, Italy, October 4-5*, 2012.

[6] H. Wang, T. Lee, C.-C. Leung, B. Ma, and H. Li. Acoustic segment modeling with spectral clustering methods. *in submission to IEEE/ASM TASLP*.

[7] H. Wang, T. Lee, C.-C. Leung, B. Ma, and H. Li. Using parallel tokenizers with DTW matrix combination for low-resource spoken term detection. In *ICASSP*, 2013.

[8] H. Wang, C.-C. Leung, T. Lee, B. Ma, and H. Li. An acoustic segment modeling approach to query-by-example spoken term detection. In *ICASSP*, 2012.

[9] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin. Speaker normalization on conversational telephone speech. In *ICASSP*, 1996.