# The CUHK Spoken Web Search System for MediaEval 2013

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 an audio keyword detection system developed at the Chinese University of Hong Kong (CUHK) for the spoken web search (SWS) task of MediaEval 2013. The system was built only on the provided unlabeled data, and each query term was represented by only one query example (from the basic set for required runs). This system was designed following the posteriorgram-based template matching framework, which used a tokenizer to convert the speech data into posteriorgrams, and then applied dynamic time warping (DTW) for keyword detection. The main features of the system are: 1) a new approach of tokenizer construction based on Gaussian component clustering (GCC) and 2) query expansion based on the technique called pitch synchronous overlap and add (PSOLA). The MTWV and ATWV of our system on the SWS2013 Evaluation set are 0.306 and 0.304.

## 1. INTRODUCTION

The spoken web search (SWS) task of MediaEval 2013 aims at detecting the keyword occurrences in a set of spoken documents using audio keyword queries in a language-independent fashion. The spoken documents involves about 20 hours of unlabeled speech data from 9 languages. More details about the task description can be found in [1]. The focus of our work was on a completely unsupervised setting, *i.e.*, only the unlabeled data from the spoken documents was used in the system development. For each query term, only one audio example was used in our system.

Our system follows the posteriorgram-based template matching framework [2]. New methods have been developed for tokenizer construction and query expansion. In addition, it was found that score normalization brought significant improvement.

# 2. SYSTEM DESCRIPTION

#### 2.1 System Overview

Fig. 1 gives the overall architecture of our system. It involves offline process and online process. The offline process (marked by the dashed window in Fig. 1) is to build the system from the spoken documents. It is divided into the stages of feature extraction, tokenizer construction, and posteriorgram generation. The offline process results in a speech tokenizer and the posteriorgrams of the spoken documents.

Copyright is held by the author/owner(s).

*MediaEval 2013 Workshop*, October 18-19, 2013, Barcelona, Spain. This research is partially supported by the CUHK MoE-Microsoft Key Laboratory and CUHK-PKU Joint Centre for Intelligence Engineering. The online process is to perform the detection task given an input query. It involves query expansion, query posteriorgram generation, DTW detection and score normalization. The query expansion is based on the PSOLA [3] technique, which modifies the duration of the original query example and generates a number of query examples of different lengths. We refer to the original query examples and the generated query examples as the *expanded query set*. After converting the *expanded query set* into posteriorgrams, DTW detection is applied to get the raw scores. DTW is performed with a sliding window on the log-inner-product distance matrix of the posteriorgrams of the query set and the spoken documents. Details of the DTW detection in our system can be found in [5]. Lastly mean and variance normalization is applied to the raw scores to obtain the final detection score.

In practice, when the query example was very short, the returned hits would contain many false alarms. A duration threshold of 0.35 second was applied to the input queries. If the duration of a query example (after silence removal) was less than the threshold, the system rejected this query example and did not return any results.

#### 2.2 Feature Extraction

Our system used 39-dimensional MFCC features. The MFCC features were processed with voice activity detection (VAD), mean and variance normalization (MVN) on the utterance level. Vocal tract length normalization (VTLN) was then used to alleviate the influence of speaker variation. The warping factors were determined with a maximum-likelihood grid search using a GMM with 256 components. The usefulness of VTLN for this task was experimentally demonstrated in our previous paper [6].

## 2.3 Tokenizer Construction

The tokenizer was used to generate posteriorgrams. It was trained from the unlabeled data of the spoken documents. We used a new Gaussian component clustering (GCC) approach to find phonemelike units, and modeled the corresponding context-dependent states by a 5-layer neural network. The posteriorgrams were composed of the state posterior probabilities produced by the neural network.

The GCC approach involved 4 steps. First, a GMM with 4096 components was estimated. Second, unsupervised phoneme segmentation was performed on the spoken documents. Third, each speech segment was represented by a Gaussian posterior vector, which is computed by averaging the frame-level Gaussian posterior probabilities. Stacking the Gaussian posterior vectors, we obtained a Gaussian-by-segment data matrix, which is denoted by X. Finally, we computed the similarity matrix W of the Gaussian components as  $W = XX^T$ , and apply spectral clustering on the similarity matrix to find 150 clusters of Gaussian components. Details of the GCC approach can be found in [4].

Each cluster of Gaussian components was viewed as the acoustic



**Figure 1: System Framework** 

model of a discovered unit. These acoustic models were refined by an iterative process [6], and updated to context-dependent models with 1198 states. These states were then modeled by a deep neural network, which had 3 hidden layers with 1000 units per layer. The input layer corresponds to a context window of 9 successive frames. The outputs of the neural network were the state posterior probabilities and used to construct the posteriorgrams.

## 2.4 Query Expansion

Query expansion aimed at generating variable length examples, so as to cover larger duration variation of the query term. The PSOLA technique was implemented for this purpose. PSOLA is able to perform time-scale modifications while preserving the spectral characteristics as much as possible. The implementation involved three steps. First, pitch epochs were detected by an autocorrelation method. Second, the periodic waveform cycles identified by the pitch marks were duplicated or eliminated according to the time-scaling factors. Finally, the overlap-and-add algorithm was used to synthesize the new speech example. In the system, two time-scaling factors were used: 0.7 and 1.3. For a query example with duration L, we had one generated example with duration  $0.7 \times L$  and another with duration  $1.3 \times L$ . Therefore the *expanded* query set would have three examples for each term. Given a query term and an utterance in the spoken documents, the detection score was the maximum value among the scores provided by the three examples.

## 2.5 Score Normalization

Let  $d_{q,t}$  denote the DTW alignment distance between the  $q_{th}$  query on the  $t_{th}$  hit region. The corresponding raw detection score was computed by

$$s_{q,t} = \exp(-d_{q,t}/\beta),\tag{1}$$

where the scaling factor  $\beta$  was set to 5. To calibrate the scores of different query terms, a simple 0/1 normalization was used. The normalization was performed as

$$\hat{s}_{q,t} = (s_{q,t} - \mu_q)/\delta_q, \qquad (2)$$

where  $\mu_q$  and  $\delta_q^2$  are the mean and variance of the top 400 raw scores for the  $q_{th}$  query.

## 3. PERFORMANCE AND ANALYSIS

Table 1 lists the performances of our systems with different configurations. System No. 3 is our submitted system for this task. All these three systems belong to the *required run* condition defined in [1]. From Table 1, we have observed severe performance degradation ( $\geq 5\%$ ) from the Dev query set to the Eval query set. This may be due to the mismatch between the Dev set and Eval set. Another observation is that the use of query expansion indeed brings improvements ( $\geq 2\%$ ) for both the Dev set and the Eval set. We think this improvement is quite encouraging. And more experiments and analysis will be done to claim the usefulness of the query expansion in the future work. The final observation is that the use of score normalization brings two considerable benefits. First, it brings about 7.7% MTWV gain on Dev set, and 7.0% on Eval set. This is different from our observation in the previous work [5]. We suspect this is related to the nonlinear transformation in (1) and the large size of the spoken documents. Second, score normalization seemed to make the decision threshold quite stable, so that the gap between MTWV and ATWV on Eval set becomes very small.

 Table 1: System Configurations and Performances.

 The basic system is without query expansion and score normalization.

	-		
System No.	1	2	3
Basic System			
Query Expansion		$\checkmark$	$\checkmark$
Score Normalization			
Dev Query Set (MTWV)	0.263	0.290	0.367
Eval Query Set (MTWV)	0.216	0.236	0.306
Dev Query Set (ATWV)	-	-	0.367
Eval Query Set (ATWV)	-	-	0.304

#### 4. HARDWARE, MEMORY, AND CPU TIME

All the experiments were performed on a computer with Intel i7-3770K CPU (3.50GHz, 4 cores), 32GB RAM and 1T hard drive. In the online process, all the posteriorgrams of the spoken documents were stored in the memory. This accelerated the online detection, but involved very high memory cost (>10GB). The computation cost in the online process was mainly caused by DTW detection. The searching speed factor of the system No.3 was about 0.018.

#### 5. **REFERENCES**

- X. Anguera, F. Metze, A. Buzo, I. Szoke, and L. Rodriguez-Fuentes. The spoken web search task. In *MediaEval 2013 Workshop*, 2013.
   T. Hazen, W. Shen, and C. White. Query-by-example
- [2] T. Hazen, W. Shen, and C. White. Query-by-example spoken term detection using phonetic posteriorgram templates. In ASRU,2009.
- [3] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, 1990.
- [4] H. Wang, T. Lee, C.C. Leung, B. Ma, and H. Li. Unsupervised mining of acoustic subword units with segment-level Gaussian posteriorgrams. In *Interspeech*, 2013.
- [5] 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.
- [6] 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.