INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Toward High-Performance Language-Independent Query-by-Example Spoken Term Detection for MediaEval 2015: Post-Evaluation Analysis

Cheung-Chi Leung[1], *Lei Wang*[1], *Haihua Xu*[2], *Jingyong Hou*[3], *Van Tung Pham*[2], *Hang Lv*[3], *Lei Xie*[3], *Xiong Xiao*[2], *Chongjia Ni*[1], *Bin Ma*[1], *Eng Siong Chng*[2], *Haizhou Li* [1,2]

[1] Institute for Infocomm Research, A*STAR, Singapore
[2] Nanyang Technological University, Singapore
[3] Northwestern Polytechnical University, Xi'an, China

ccleung@i2r.a-star.edu.sg

Abstract

This paper documents the significant components of a state-of-the-art language-independent query-by-example spoken term detection system designed for the Query by Example Search on Speech Task (QUESST) in MediaEval 2015. We developed exact and partial matching DTW systems, and WFST based symbolic search systems to handle different types of search queries. To handle the noisy and reverberant speech in the task, we trained tokenizers using data augmented with different noise and reverberation conditions. Our post-evaluation analysis showed that the phone boundary label provided by the improved tokenizers brings more accurate speech activity detection in DTW systems. We argue that acoustic condition mismatch is possibly a more important factor than language mismatch for obtaining consistent gain from stacked bottleneck features. Our post-evaluation system, involving a smaller number of component systems, can outperform our submitted systems, which performed the best for the task.

Index Terms: Data augmentation, bottleneck features, dynamic time warping, symbolic search, partial matching

1. Introduction

Query-by-example spoken term detection (QbE-STD) refers to the task of finding the occurrences of a spoken query in an audio archive. This task does not necessarily require the linguistic knowledge and transcribed data of the target language, and it has gained the interest of the research community in recent years. The query-by-example search on speech task (QUESST) (and formerly the spoken web search task), which has been held in recent MediaEval evaluation campaigns and provided suitable benchmarks for the development of QbE-STD, is also one of the driving forces of this research direction.

The QUESST 2015 dataset [1], similar to last year, consists of speech from heterogeneous sources in multiple languages. The major challenge of the QUESST 2015 is the presence of noise and reverberations in the dataset. The popular approach to QbE-STD is acoustic pattern matching based on variants of dynamic time warping (DTW) [2, 3]. Posterior features or bottleneck features from neural networks trained using mismatched languages, and fusion of multiple systems are usually used in top-performing systems [4-10]. Unsupervised acoustic modeling or feature extraction has been studied in [11-14] to deal with the lack of knowledge about target data. Partial matching techniques [15-17] have been developed to deal with different kinds of query matches for the QUESST 2014.

This paper summarizes the significant components of our system for the QUESST 2015, and present the post-evaluation effort devoted to improve our system. To address the noise and reverberations in the QUESST 2015 data, training tokenizers with data augmentation was attempted in our submitted system [10]. Its aim is to make the tokenizers to have better coverage of various acoustic conditions in the QUESST data. Our post-evaluation effort was firstly devoted to training of tokenizers with better data augmentation, and then we revisited some major techniques, including speech activity detection (SAD) and the search backends, with the features from the improved tokenizers. The better data augmentation was achieved by augmenting the training data with different types of noises from the NOISEX-92 corpus [18] and reverberations synthesized by room impulse response.

In the post-evaluation analysis, when the tokenizers were trained with the improved data augmentation method, we investigated its interaction with a number of major components, and a number of observations are highlighted as follows.

- The phone boundary label provided by the improved DNN tokenizers brings more accurate speech activity detection in DTW systems, and this eliminates the use of multiple detectors in our submitted system.
- The bottleneck features (referred to as stacked bottleneck features) extracted from the second-level network of a stacked hierarchical neural network [19] can consistently outperform the bottleneck features from the first-level network; Although it is not surprising in speech recognition with an expected test condition [20], it is, to our best knowledge, the first reported result in QbE-STD on multiple languages. In the experiments on other data, we observe that the longer temporal context brought by stacked bottleneck features do not necessarily bring improvement. We argue that acoustic condition mismatch is possibly a more important factor than language mismatch for obtaining consistent gain from stacked bottleneck features.
- The post-evaluation system with four component systems outperforms our submitted system, which consisted of 66 component systems and performed the best among all the QUESST 2015 teams.

http://dx.doi.org/10.21437/Interspeech.2016-691

2. Evaluation corpus and task

The QUESST 2015 dataset is collected from multiple languages (including Albanian, Czech, Singaporean Mandarin Chinese mixed with Singaporean English, Portuguese, Romanian and Slovak) and sources with different recording environments and speaking styles. It consists of around 18 hours of audio, 445 development (dev) queries and 447 evaluation (eval) queries. Noise (collected from https://www.freesound.org) and reverberations are artificially added to the dataset. For details, please refer to the QUESST 2015's evaluation specification [1].

The task requires the detection systems to return the following three types of matches: 1) exact match (T1) in which the query is read speech; 2) partial match (T2) which allows reordering and small lexical variation between query and test utterances, and in which the query is dictated speech; 3) partial match (T3) which allows reordering and small lexical variation between query and test utterances, and in which the query is conversational speech. For each query, no prior information regarding the spoken language or the type of matches involved is available to the detection systems.

System performance is evaluated using normalized cross entropy (Cnxe) and term weighted value (TWV). Smaller values of Cnxe and larger values of TWV indicate better performance.

3. Overview of our submitted system

Our submitted system involves dynamic time warping (DTW) and symbolic search based backends similar to our submitted system [8] for QUESST 2014. However, to address the more challenging acoustic and noise conditions of the data in QUESST 2015, we attempted to estimate noise in the data using the techniques in [21-23], and added the noise to a portion of training data of some tokenizers. We also used a large number of tokenizers trained using different speech corpora, two speech activity detectors. The DNN models which were used as bottleneck feature extractors and phone recognizers were trained using the Kaldi toolkit. For detail, please refer to our system description [10].

3.1. DTW systems

Bottleneck features (BNF), stacked bottleneck features (SBNF) and phoneme-state posterior features were mainly used in our DTW systems. Exact matching and partial matching DTW systems were developed to deal with different types of queries. We used two speech activity detectors, including frequency band energy based SAD [23] and statistical model based SAD [24], to remove non-speech frames in utterances. It was because we found that they performed the best in different types of queries in DTW systems.

An exact matching system matched each query with a subsequence of a test utterance using subsequence DTW [3]. It found a path on the cosine distance matrix of the speech features of the query and the test utterance. The system output the similarity score between the query and the matched subsequence of the test utterance.

Our partial matching DTW systems, including fixed-window [8, 16] and phoneme-sequence [17] partial matching systems, were used to deal with T2 and T3 queries. In each fixed-window partial matching system, an analysis window between 70 and 90 frames long was defined. When the window was shifted between 5 and 10 frames in each step, a query segment from the analysis window was matched with a test utterance. The highest similarity score which corresponded to a query segment and the test utterance was used as the score of the query-utterance pair of the system. In phoneme-sequence approximate matching systems, the size of the window was determined by the phoneme boundary information derived from phoneme recognizers. The window size was set to 8 phonemes, as it provided best results on the development data.

3.2. WFST-based symbolic search systems

Weighted finite state transducer (WFST) based symbolic search systems were used to deal with T2 and T3 queries [8, 16]. Such systems decoded a query utterance into N-best phone sequences, and the partial phone sequences were extracted and converted to WFST format. The phoneme lattice of search utterances was converted into timed factor transducer [27]. The search was performed by the composition of query and search audio WFSTs. Although symbolic search systems do not perform as good as DTW systems for partial matching, it allows indexing of the search utterances and facilities fast search.

4. Post-evaluation system and analysis

In post-evaluation analysis, we kept the overall architecture of our system, which is based on DTW search and WSFT symbolic search, while the data augmentation method was revised and the total number of component systems was greatly reduced. When the revised data augmentation method was used, we observed that the tokenizers could give better SAD using its phone-level decoding, and this eliminated multiple component systems using different SAD in our submitted system.

Only SBNFs were used in the DTW systems of our post-evaluation system. We trained SBNF extractors using stacked hierarchical networks. The first-stage network took filterbank and pitch features as input. The first-stage and the second-stage networks had the topology of 1500-1500-80-1500-x, and 1500-1500-40-1500-x respectively, where x is the number of senones (around 400 in different tokenizers).

We also used fewer speech corpora, including Switchboard English (LDC97S62) and Fisher Spanish (LDC LDC2010S01), to train tokenizers; they performed better than tokenizers trained using other corpora. We performed the fusion of 4 component systems to achieve the performance that is better than our submitted system for the evaluation.

4.1. Data augmentation

In our submitted system, we extracted the noise segments in the evaluation data and augmented them to the data for training a tokenizer. Although performance improvement (4% relative improvement in exact matching system for T1 dev queries) was observed by this method [10], we noted that the effect of reverberations was not considered and the newly trained tokenizer was only targeted to the noise condition in the QUESST 2015.

To address these issues, we augmented the original training data with different reverberation and noise conditions. Firstly we convolved the original training data of the tokenizers with room impulse responses, which were

artificially generated using the image model [28]. Three room sizes, including small, medium and large, were considered. Two speaker-to-microphone distances, including 1.5 meters and 3 meters, were used. Reverberation time T60 was randomly chosen between 0.1 sec and 1.0 sec. After an original training utterance was convolved with a RIR, additive noise with signal-to-noise ratio (SNR) randomly selected between 0dB and 50dB was added. Totally 15 types of additive noises samples from the NOISEX-92 database [18] were used to contaminate the utterance. For each original utterance, its RIR and noise sample were randomly selected. Finally each tokenizer was trained using the double amount of training data (the original data and the data contaminated with different reverberation and noise conditions).

We also observed that when a phone recognizer was trained with data augmentation, the phone recognizer could bring more accurate phone boundary for speech activity detection. It was used to remove non-speech frames in query utterances for DTW search because we expected the non-speech frames do not help the search of spoken words in the query utterances.

Table 1 shows the improvement of an exact matching DTW system, in which the Switchboard SBNF extractor was used and it was trained with the revised data augmentation procedures. When the spectral energy based SAD is used to drop non-speech frames, the SBNF can bring 10% relative improvement (from 0.758 to 0.684) in minCnxe for T1 queries. If a phone recognizer trained using the Switchboard corpus with data augmentation is used to perform SAD, we can further obtain 9% relative improvement (to 0.621) in minCnxe. We also observed similar improvements in other tokenizers and partial matching DTW systems [17].

Table 1. *Effect of SAD and data augmentation on QUESST 2015 dev set.*

SAD	With data augmentation?	minCnxe/maxTWV	
		T1 queries	all queries
No	No	0.812 / 0.158	0.909 / 0.084
Spectral energy based	No	0.758 / 0.234	0.889 / 0.114
Spectral energy based	Yes in SBNF	0.684 / 0.326	0.847 / 0.167
Time label from phoneme recognizer	Yes in SBNF and SAD	**0.621** / **0.412**	**0.827** / **0.214**

To ensure whether the data augmentation hurt the clean spoken queries, we further analyzed the improvement for clean utterances and all utterances. Surprisingly, we found that improvement obtained in clean T1 utterances is comparable with the improvement in T1 queries. Perhaps the "clean" utterances (without artificial noise and reverberations) originally contains certain noise or/and reverberations.

Table 2. *Effect of data augmentation on QUESST 2014 dev set. Exact matching DTW system with Switchboard SBNF is used. No artificially noise and reverberations in this dataset.*

With data augmentation?	minCnxe/maxTWV	
	T1 queries	all queries
No	0.747 / 0.309	0.854 / 0.183
Yes	**0.740** / **0.312**	**0.851** / **0.187**

Moreover, we also used the Switchboard tokenizer with data augmentation to run the experiments using the QUESST 2014 dataset, and we observed no performance hurt at the QUESST 2014 dataset as shown in Table 2. We believe that the resultant tokenizer has better coverage of various acoustic conditions.

4.2. Improvement by stacked bottleneck features and effect of data mismatch

Stacked bottleneck features (SBNF) [19], which are extracted from the second-level network and make use of longer temporal context, can outperform the bottleneck features (BNF) from the first-level neural network in ASR. However, it is usually not the situation in QbE-STD on multiple languages.

In our submitted system, we used both BNF and SBNF. It was motivated by the observation (as shown in Table 3) that SBNF could not bring consistent gain over BNF across queries in different languages when no data augmentation was used. Previously, we attributed it to the language mismatch between the SBNF tokenizer and the audio archive. However, we found that when the data augmentation method was used, SBNF consistently outperformed BNF, and the fusion with the system using the bottleneck features from the first-level neural network was not necessary. To our knowledge, consistently superior performance by SBNF has never been reported in QbE-STD on multiple languages.

We also used the Switchboard BNF and SBNF extractors (without data augmentation) to run the QbE-STD experiments on the QUESST 2014 dataset (similarly from different sources, but with less acoustic variation), we also observed that SBNF only brought performance gain on English (non-native) queries as shown in Table 4. Although there is no overall performance drop in all non-English (including Albanian, Basque Czech, Romanian and Slovak) queries, obvious performance drop is found in Albanian and Basque queries.

In a word-discrimination task on the Switchboard corpus reported in [29], cross-lingual BNF and SBNF extractors were trained using similar telephony conversational speech, but from different languages (Spanish and Mandarin Chinese). In the experiments, cross-lingual SBNF extractors trained using the Fisher Spanish and HKUST Mandarin Chinese corpora could outperform their corresponding BNF extractors.

Based on the observations from the above three different sets of experiments, we believe that reducing data mismatch (especially mismatch in language and acoustic variation) is an important factor for obtaining consistent gain from cross-lingual SBNF on multi-lingual QbE-STD. In the acoustic challenging audio like the QUESST 2015 dataset, reducing the acoustic condition mismatch becomes more important.

Table 3. *SBNF outperform BNF when data augmentation is used. Tokenizers are trained using Switchboard English corpus. Results are evaluated on T1 dev queries of QUESST 2015.*

With data augmentation?	Features	minCnxe / maxTWV						
		Albanian	Mandarin	Czech	Portugese	Romanian	Slovak	All
No	BNF	0.468 / 0.433	0.951 / 0.057	0.838 / 0.116	0.561 / 0.394	0.433 / 0.559	0.645 / 0.229	0.744 / 0.240
	SBNF	0.572 / 0.381	0.945 / 0.075	0.827 / 0.148	0.579 / 0.367	0.457 / 0.542	0.659 / 0.258	0.758 / 0.234
Yes	BNF	0.417 / 0.468	0.911 / 0.090	0.771 / 0.210	0.454 / 0.473	0.423 / 0.564	0.535 / 0.432	0.680 / 0.330
	SBNF	**0.399** / **0.572**	**0.891** / **0.138**	**0.711** / **0.251**	**0.349** / **0.626**	**0.362** / **0.667**	**0.505** / **0.477**	**0.621** / **0.412**

Table 4. *Performance of BNF and SBNF on QUESST 2014 dataset. Results are evaluated on T1 dev queries.*

Tokenizers	minCnxe / maxTWV	
	English	non-English
Switchboard BNF	0.865 / 0.092	0.746 / 0.335
Switchboard SBNF	**0.816** / **0.146**	**0.742** / **0.336**

4.3. Partial matching

When the revised data augmentation method was used, both partial matching DTW systems and symbolic search systems could obtain considerable performance gain. An English phoneme tokenizer, which was trained using the Switchboard corpus, was used in a partial matching DTW (for obtaining phoneme boundary) and a symbolic search system. The phoneme tokenizer was a 6-hidden-layer DNN with 2048 neuron units in each layer, and it used a phoneme-loop grammar to perform tokenization. The partial matching DTW system using the phoneme boundary information showed slightly better performance than that using fixed window. For more detailed analysis of the partial matching systems, please see [17].

Table 5. *Performance of three post-evaluation component systems on QUESST 2015 dev set. All use tokenizers trained using Switchboard.*

Systems	minCnxe / maxTWV			
	T1	T2	T3	All
Exact matching DTW	**0.621** / **0.412**	0.886 / 0.090	0.880 / 0.132	0.827 / 0.214
Partial matching DTW	0.701 / 0.329	**0.781** / **0.193**	**0.838** / **0.149**	**0.788** / **0.225**
Symbolic search	0.903 / 0.108	0.925 / 0.029	0.949 / 0.031	0.931 / 0.055

The performance of the phoneme-boundary partial matching DTW system, the symbolic search system and the exact matching system mentioned in section 4.1 are shown in Table 5. From the table, we can observe the partial matching DTW system performs better than for T2 and T3 queries, while the exact matching system performs better for T1 queries. Although the symbolic search system is not as good as the other two systems, we will show that these systems complement each other in system fusion in the next section.

4.4. System fusion

As in our submitted system, scores from DTW systems were normalized to zero mean and unit variance, and scores from symbolic search systems were converted to log-likelihood ratio. Scores from all component systems were then fused with the FoCal toolkit [30].

Table 6. *Performance of fused systems on QUESST 2015 dev set.*

Systems	minCnxe / maxTWV			
	T1	T2	T3	all
Exact matching DTW (English) + Exact matching DTW (Spanish)	0.566 / 0.466	0.863 / 0.128	0.852 / 0.170	0.795 / 0.256
Partial matching DTW + Symbolic search	0.699 / 0.344	0.777 / 0.192	0.832 / 0.155	0.783 / 0.231
Fusion of four systems	**0.558** / **0.480**	**0.753** / **0.259**	**0.784** / **0.219**	**0.723** / **0.320**

Table 6 shows the results of the fused systems in post-evaluation. One more exact matching DTW system using the Fisher Spanish SBNF with data augmentation is involved in the system fusion. It uses the same phoneme-label-based SAD as the other exact matching DTW system. Performance gains are obtained when fusing exact matching and two types of approximate matching systems, and fusing systems using different tokenizers. Note that fusion of the four systems (minCnxe of 0.723 and MTWV of 0.320) outperforms our submitted system (minCnxe of 0.757 and MTWV of 0.286) which performed the best among all the QUESST 2015 teams.

5. Conclusions

The significant techniques for building a state-of-the-art language-independent QbE-STD system for the QUESST 2015 have been summarized. Our post-evaluation system involving a smaller number of component systems outperforms our submitted system, which performed the best for the task. To deal with the challenging acoustic conditions of the data, training of tokenizers with data augmentation has been shown important in different types of component systems. Score fusion of different component systems is still important for obtaining considerable performance gain.

6. References

[1] I. Szoke, L. J. Rodriguez-Fuentes, A. Buzo, X. Anguera, F. Metze, J. Proenca, M. Lojka, and X. Xiao, "Query by Example Search on Speech at MediaEval 2015," in *Proc. MediaEval workshop*, 2015.

[2] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.

[3] A. Muscariello, G. Gravier, and F. Bimbot, "Audio keyword extraction by unsupervised word discovery," in *Proc. INTERSPEECH*, 2009, pp. 2843-2846.

[4] T. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Proc. IEEE ASRU*, 2009, pp. 421–426.

[5] J. Tejedor, M. Fapšo, I. Szöke, J. Černocký, F. Grézl, "Comparison of methods for language-dependent and language-independent query-by-example spoken term detection," *ACM Transactions on Information Systems (TOIS)*, vol. 30, no. 3, 2012.

[6] 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 *Proc. ICASSP*, 2013, pp. 8545-8549.

[7] L. J. Rodriguez-Fuentes, A. Varona, M. Penagarikano, G. Bordel, and M. Diez, "High-performance query-by-example spoken term detection on the SWS 2013 evaluation," in *Proc. ICASSP*, 2014, pp. 7819-7823.

[8] P. Yang *et al*, "The NNI query-by-example system for MediaEval 2014," in *Proc. MediaEval workshop*, 2014.

[9] I. Szöke, M. Skácel, L. Lurget, J. Černocký, "Coping with channel mismatch in Query-by-Example - BUT QUESST 2014," in *Proc. ICASSP*, 2015, pp. 5838-5842.

[10] J. Hou *et al*, "The NNI query-by-example system for MediaEval 2015," in *Proc. MediaEval workshop*, 2015.

[11] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *Proc. IEEE ASRU*, 2009, pp. 398–403.

[12] H. Wang, T. Lee, C.-C. Leung, B. Ma, and H. Li, "Acoustic segment modeling with spectral clustering methods," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, 264-277, 2015.

[13] P. Yang, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Intrinsic spectral analysis based on temporal context features for query-by-example spoken term detection," in *Proc. INTERSPEECH*, 2014, pp. 1722-1726.

[14] H. Chen, C.-C. Leung, L. Xie, B. Ma, H. Li, "Unsupervised bottleneck features for low-resource query-by-example spoken term detection," in *Proc. INTERSPEECH*, 2016.

[15] J. Proença, A. Veiga, and F. Perdigão, "The SPL-IT-UC Query by Example Search on Speech System for MediaEval 2014," in *Proc. MediaEval workshop*, 2014.

[16] H. Xu *et al*, "Language independent query-by-example spoken term detection using n-best phone sequences and partial matching," in *Proc. ICASSP*, 2015, pp. 5191-5195.

[17] H. Xu *et al*, "Approximate search of audio queries by using DTW with phone time boundary and data augmentation," in *Proc.* ICASSP, 2016, pp. 6030-6034.

[18] A. Varga, H.J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, 12(3), pp. 247-251, 1993.

[19] F. Grezl and M. Martin, "Hierarchical neural net architectures for feature extraction in ASR," in *Proc. INTERSPEECH*, 2010, pp. 1201-1204.

[20] C. Plahl, R. Schlüter, H. Ney, "Cross-lingual portability of Chinese and English neural network features for French and German LVCSR," in *Proc. IEEE ASRU*, 2011, pp. 371-376.

[21] W. Yao and T. Yao, "Analyzing classical spectral estimation by MATLAB," *Journal of Huazhong University of Science and Technology*, vol. 4, p. 021, 2000.

[22] M. H. Hayes, J. S. Lim, and A. V. Oppenheim, "Signal reconstruction from phase or magnitude," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 6, pp. 672–680, 1980.

[23] M. H. Gruber, "Statistical digital signal processing and modeling," *Technometrics*, vol. 39, no. 3, pp. 335–336, 1997.

[24] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1218-1234, 2006.

[25] E. Cornu, H. Sheikhzadeh, R. L. Brennan, H. R. Abutalebi, E. C. Tam, P. Iles, and K. W. Wong, "ETSI AMR-2 VAD: evaluation and ultra low-resource implementation," in *Proc. ICME*, 2003, pp. II–841-4.

[26] M. Huijbregts and F. De Jong, "Robust speech/non-speech classification in heterogeneous multimedia content," *Speech Communication*, vol. 53, no. 2, pp. 143–153, 2011.

[27] D. Can and M. Saraclar, "Lattice indexing for spoken term detection," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 2338–2347, 2011.

[28] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[29] Y. Yuan, C.-C. Leung, L. Xie, B. Ma, H. Li, "Learning neural network representations using cross-lingual bottleneck features with word-pair information," in *Proc. INTERSPEECH*, 2016.

[30] N. Brummer, "Focal toolkit," available online: https://sites.google.com/site/nikobrummer/focal.