EFFECTIVE UTILIZATION OF MULTIPLE EXAMPLES IN QUERY-BY-EXAMPLE SPOKEN TERM DETECTION

Ji Xu, Ge Zhang, Yonghong Yan

The Key Laboratory of Speech Acoustics and Content Understanding,
Chinese Academy of Sciences
Beijing 100190, P.R.China
{xuji, zhangge, yanyonghong}@hccl.ioa.ac.cn

ABSTRACT

This paper investigates the example utilization problem in query-by-example spoken term detection when multiple examples are provided for each query term. To achieve this goal, we propose three evaluation metrics to assess the quality of all the examples, namely posteriorgram stability score, pronunciation reliability score and local similarity score. We also present a clustering based example generation approach to creating better examples based on the original ones. Experiments conducted on a telephone speech corpus shows that it is better to use several representative examples selected by the quality assessment process than to simply use all the examples. Furthermore, even better results can be obtained if the generated examples are used.

Index Terms— spoken term detection, query-by-example, multiple examples utilization, example quality assessment, example generation

1. INTRODUCTION

In recent years, state-of-the-art spoken term detection (STD) systems are usually based on large vocabulary continuous speech recognition approaches (LVCSR) [1-3]. In such a framework, speech utterances are first converted into lattices, and a text based search is then applied. LVCSR based STD systems provide satisfying performance in well-resourced languages [4-7], but it is difficult to build such systems when there is no enough data for statistical acoustic and language model training, which is common for several applications.

Query-by-example spoken term detection (QbE-STD) is a task in which queries are given in utterance examples, instead of orthographic forms. This provides an opportunity to avoid resource-consuming language specific statistical modeling by searching queries in the speech documents directly based on their signal characteristics [8-15]. In a typical template-matching based framework of QbE-STD, speech documents along with the queries are firstly represented by posteriorgrams, like phoneme posteriorgrams [8], gaussian posteriorgrams [9] or acoustic segment model (ASM) posteriorgrams [12]. A dynamic time warping (DTW) based approach is then applied to search for the best matches between them.

This paper focuses on example utilization in template-matching based QbE-STD when multiple examples of a query term are available, especially in the case where examples are spoken by various speakers with different styles. It is common in practical scenarios as the examples provided by users always have personal characteristics, e.g. accent. Therefore, using all the examples without distinction may not be optimal.

Currently, there are two common approaches for multiple examples utilization. The first one is to create a single average example by aligning all other examples to a base example, then using it to search for queries just like single example condition [16]. This approach is computationally cheaper but it relies heavily on the quality of the base example [17]. The second one is to use all available examples to generate scores independently, and a simple average or score fusion strategy is then performed to obtain the final score of the candidate region [8, 9, 18]. The advantage of this approach is that the candidate score is more stable than the first approach, as it takes into account the contribution of all examples in parallel. As a trade-off, a lot of calculations are needed in this approach, which may be demanding in some practical applications.

To overcome the disadvantages in the above approaches, this paper presents a novel approach of multiple examples utilization in two steps, example quality assessment and example generation. In the first step, evaluation metrics are used to assess the examples on stability, reliability and similarity, and several representative examples are selected based on their ranks. In the second step, the posteriorgrams of the selected examples are used as seeds and new examples are generated by adjusting the posteriorgrams according to the examples in the same cluster.

This paper is organized as follows. Section 2 describes the method of example quality assessment. Section 3 introduces the example generation approach. Experiments along with a discussion are presented in Section 4 and conclusion is drawn finally in Section 5.

2. EXAMPLE QUALITY ASSESSMENT

It is obvious that there is no single example that can outperform all other examples in all situations, but some rules can be used to predict which examples may get better results in general. We tackle this problem from three respects, just similar to the way humans evaluate acoustic examples. The first one is whether the pronunciation of an example is "clear", which means a good example should be consist of a sequence of basic acoustic units clearly. The second one is whether the pronunciation of an example is "correct", which means that the basic acoustic units used by a good example should be the commonly used ones. The third one is whether an example is "similar" to other examples in general, which aims at avoiding some extreme cases like inappropriate speaking rate. Therefore, three evaluation metrics are proposed, namely posteriorgram stability score (PS score), pronunciation reliability score (PR score) and local similarity score (LS score).

2.1. Posteriorgram stability score

After the posteriorgram of an example is generated, we employ hierarchical agglomerative clustering (HAC) algorithm [19] to convert it into a sequence of segments. Given a posteriorgram $\mathrm{P} = \{P_1, \cdots, P_n\}$, segments $\mathrm{S} = \{S_1, \cdots, S_m\}$ is initialized with $\mathrm{m} = \mathrm{n}$. and $S_i = P_i$. In each step, two adjacent segments are merged according to the criterion that minimizes the following function:

$$\mathrm{Q(P,S)} = \sum_{i=1}^{S} \sum_{j=b_i}^{e_i} d\left(P_j, c_i\right) \quad (1)$$

where c_i represents the centroid of the i^{th} segment and $d\left(P_j, c_i\right)$ is the distance between P_j and c_i. b_i and e_i are the beginning and ending frame number of the i^{th} segment. The merging process will not stop until the increment of $\mathrm{Q(P,S)}$ reaches a certain threshold δ, which is manually set. If δ is very large, only a single segment will be left at the end. We control the threshold to zoom the segments to the phoneme level.

Considering the short-term stability of speech signals, if a query is clearly spoken and successfully characterized by the front-end classifier, the posteriorgram of each segment generated from HAC should be relatively stable, especially for the acoustic units with higher probability. Thus, we use the average probability of the top N acoustic units as PS score, which can avoid being disturbed by the minor ones within a segment. PS score is defined as follows:

$$\mathrm{PS} = \sum_{\mathrm{i=1}}^{\mathrm{S}} \sum_{\mathrm{j=b_i}}^{\mathrm{e_i}} \sum_{\mathrm{n=1}}^{\mathrm{N}} \mathrm{P_{jn}} \quad (2)$$

where P_{jn} is the probability of the n^{th} top acoustic units in j^{th} frame. The top acoustic units are selected at segment level.

The PS score describes whether the posteriorgram of an example is stabilized in general. Although we cannot guarantee that an example with high PS score can obtain good result, the example with low PS score always results in high false alarm rate as the lack of language level constraint in template-matching based framework. Therefore, it is reasonable to use PS score as one of the evaluation metrics.

2.2. Pronunciation reliability score

To calculate the PR score, we first represent the examples as a sequence of acoustic units, which are the ones with the highest probabilities in the segments generated by HAC. In the next step, Levenshtein distance between every two examples with the same term is calculated, preserving the detailed information of substitution, insertion and deletion. For two examples q_i and q_j, we define a reliability contribution score $\mathrm{c}(q_i, q_j)$ as follows:

$$\mathrm{c\left(q_i, q_j\right)} = \max(1 - \mathrm{aN_{sub}} - \mathrm{bN_{ins/del}}, 0) \quad (3)$$

where N_{sub} is the number of substitutions and $N_{ins/del}$ is the sum of insertions and deletions. a and b are two parameters which are related to the average segments number of a query term. As an example, we use $a = 0.3$ and $b = 0.4$ if the average segment number is lower than 10. We make $a < b$ because we consider that misclassification may occur when the training data and the test data are not exactly matched. The PR score is defined as:

$$\mathrm{PR(q_i)} = \sum_{\mathrm{j \neq i}} \mathrm{c\left(q_i, q_j\right)} \quad (4)$$

The PR score describes if an example is close to other examples in pronunciation. An example with abnormal pronunciation will not become a recommended one.

2.3. Local similarity score

For a certain example, we calculate the DTW distances between this example and all examples with the same term. The distances are ranked and K lowest values are selected. The LS score is defined as the mean value of the selected distances, which can be expressed as follows:

$$\mathrm{LS(q_i)} = \frac{1}{K} \sum_{j^*=1}^{K} \mathrm{DTW\left(q_i, q_{j^*}\right)} \quad (5)$$

where q_{j^*} is the j^{th} nearest example for $\mathrm{q_i}$.

The LS score describes whether an example is similar to "nearby" examples. We do not use the mean value of all the distances since the examples are diverse. Therefore, a global mean value may not provide useful information.

2.4. Selection of representative pronunciations

Considering that many spoken terms can have multiple pronunciations, using a single example may not be the best choice. Hence, several representative pronunciations are selected based on their ranks in quality assessment. In order not to select examples that are too similar, an example will be skipped if its DTW distance with an already selected example is smaller than a certain threshold.

3. CLUSTERING BASED EXAMPLE GENERATION

This section aims at creating better examples from the given ones. First we select several original examples as seeds through example quality assessment. For each seed, the following schedule is adopted to improve its quality, which is similar to the framework of k-means clustering algorithm:

Step 1: Set the current example as the seed.

Step 2: Calculate the DTW distance between the current example and all examples with the same term. Select K examples with the lowest scores, which are the "nearest" examples. If maximum iteration number is reached or the K examples are the same as the previous iteration, go to step 6.

Step 3: Calculate the mean value of the DTW distance of the selected examples, i.e. the LS score. Set an initial learning rate λ.

Step 4: For each acoustic unit of each frame in the current example, increase its posterior by λ and decrease the other posteriors in the same frame by λ in total. The decrement is proportional to the current value of those posteriors. Save the modified example as a candidate.

Step 5: Calculate the LS scores of all candidates and find the one with the minimum LS score. If the decrement in LS score is greater than a predefined threshold, replace the current example by the candidate and go to step 4. Otherwise, half the learning rate λ and go to step 6.

Step 6: If the learning rate λ is greater than a predefined threshold, go to step 4. Otherwise, go to step 2 for the next iteration.

Step 7: Save the current example and exit.

For the three evaluation metrics used in example quality assessment, the LS score is the optimization objective of the algorithm and will be improved in the generated example; the PS score will increase in most cases as the algorithm is seeking for the commonalities, and the differences in minor acoustic units will be removed; the PR score will not be greatly changed in general because examples in the same cluster always have similar pronunciations. Therefore, the examples generated by the algorithm can usually get higher score than the seed, and may become better candidates for template matching.

4. EXPERIMENTS

4.1. Experimental setup

To simulate a specific application scenario, we use an evaluation set constructed by our own. In this application, only queries are given during development phase, which is a little different from QUESST evaluation [20, 21]. During testing phase, the pre-given queries are searched in an evaluation data set, no additional information about queries is provided.

In our experiments, we select 15 Chinese words as queries, each with 60 examples. All the examples are automatically cut from a Mandarin telephone speech corpus. We have examined all the examples to ensure that they can be recognized by humans, but do not restrict their speaking styles. It is a common case in real-life applications because the users do not always provide standard examples. The evaluation set includes 9.8 hours of Mandarin telephone speech, which does not include the given examples.

Two performance metrics are used in this paper. The first one is the average precision of the top N hits (P@N), where N is number of occurrences of the term in the test set [8]. The second one is the F-measure, which is commonly used in the field of information retrieval.

4.2. Baseline system

For the baseline system, phoneme posteriorgrams are first generated using a neural network based phoneme classifier, which is trained with Mandarin telephone speech. For each frame, if a non-speech posterior gets the highest value, we remove it from the posteriorgram [16]. After that, the posteriorgrams of speech documents and queries are compared using subsequence DTW [22]. The scores of the candidates are finally normalized using m-norm [13]. Three different methods of using multiple examples are used as baselines: 1) select a random example among all the examples; 2) align all the examples to the longest example [16]; 3) use all the examples independently and average their scores [8].

The results of the baseline systems are listed in the first three lines in Table 1. We can see that the score averaging method achieved the best performance as expected.

4.3. Quality assessment experiments

The lower part of Table 1 shows the results of example quality assessment. We first evaluated the three different metrics by using them separately. The results shows that using the 1st rank example selected by PR score got the best result, which means that pronunciation reliability is the most important respect of an example. The PS score got the worst result, but still much higher than the result of random selection because fewer false alarms were generated by examples with higher PS score.

The result of combining all three metrics is shown in line (7). All of the three metrics were normalized, with the best score equal to 1 and the worst score equal to 0. The result of line (7) is better than line (4) to line (6). It confirms that the combined evaluation is effective.

We selected several representative pronunciations as Section 2.4, and used them in the way of score averaging.

The results are listed in line (8) to line (10). It can be seen that using three pronunciations was much better than using a single one, but the performance decreased when five or more pronunciations were used. The reason is that there are a limited number of effective pronunciations, adding extra pronunciations may not provide useful information.

Table 1. Experiment results of example quality assessment

	Method	P@N	F-measure
(1)	Random selection	17.91%	0.1987
(2)	Aligning to the longest	26.65%	0.2405
(3)	Score averaging	29.21%	0.2705
(4)	PS score only	26.65%	0.2704
(5)	PR score only	35.39%	0.3096
(6)	LS score only	33.68%	0.2979
(7)	Combined assessment	37.74%	0.3508
(8)	(7)+3 Pron	43.07%	0.3901
(9)	(7)+5 Pron	43.07%	0.3786
(10)	(7)+7 Pron	42.22%	0.3682

4.4. Example generation experiments

In the experiments of example generation, we controlled the parameter K used in the algorithm. The results are listed in Table 2. All the experiments used the same seeds which were obtained from the previous experiment with the best result, i.e. line (8) in Table 1. From the results, we can see that the best result was obtained at $K = 6$, which was 10% of all the examples. Besides, more than 5% and 7% absolute improvements were obtained in P@N and F-measure respectively, which confirms the effectiveness of the new algorithm.

Additionally, we present posteriorgrams of four different examples with the same term, which are (a) random selection; (b) aligning to the longest; (c) 1st rank example using combined assessment; (d) the generated example using (c) as seed. The horizon axis represents the frame number and the vertical axis represents the individual phone classes. It can be seen that the posteriorgram of (a) is scattered, with a normalized PS score of 0.437 and a normalized PR score of 0.042. Therefore, it was not a recommended example and we also got a poor performance of P@N 0.174. In contrast, (d) has a normalized PS score of 0.945 and a normalized PR score of 1. A better result of P@N 0.521 was obtained.

Table 2. Experiment results of example generation

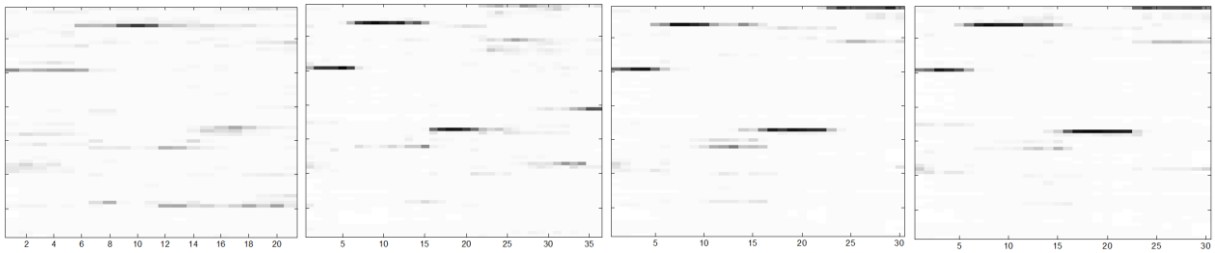
K	P@N	F-measure
3	48.19%	0.4489
6	48.40%	0.4610
9	47.33%	0.4382
12	46.70%	0.4264
18	46.06%	0.4225
60	45.84%	0.4100

5. CONCLUSION

A novel approach of multiple examples utilization for query-by-example spoken term detection is proposed in this paper. In this approach, examples are firstly assessed by three metrics in stability, reliability and similarity. A clustering based example generation algorithm is then applied to the selected examples and new examples are created. Our experiments have shown that better performance can be obtained by using examples with higher ranks in the proposed quality assessment method, compared with conventional multiple examples utilization method. Besides, the generated examples outperform the original ones, which demonstrates the effectiveness of the new algorithm.

6. ACKNOWLEDGMENT

This work is partially supported by the National Natural Science Foundation of China (Nos. 11161140319, 91120001, 61271426), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100, XDA06030500), the National 863 Program (No. 2012AA012503) and the CAS Priority Deployment Project (No. KGZD-EW-103-2).



(a) random selection

(b) aligning to the longest

(c) 1st rank example using combined assessment

(d) the generated example using (c) as seed

Fig. 1. Posteriorgrams of different examples

7. REFERENCES

[1] M. Saraclar, R. Sproat, “Lattice-based search for spoken utterance retrieval,” in *Proc. HLT 2004*, Boston, Massachusetts, USA, pp.129-136, 2004.

[2] M. Zhou, P. Yu, C. Chelba, F. Seide, “Towards spoken document retrieval for the Internet: Lattice indexing for large-scale web-search architectures,” in *Proc. HLT’2006*, New York, 2006.

[3] J. Mamou, B. Ramabhadran, and O. Siohan, “Vocabulary independent spoken term detection,” in *Proc. SIGIR*, 2007.

[4] D.R.H. Miller, et al., "Rapid and accurate spoken term detection," in *Proc. Interspeech*, pp. 314-317, 2007.

[5] D. Can et al., “Effect of pronounciations on OOV queries in spoken term detection,” *in Proc. ICASSP*, 2009.

[6] C. Parada, A. Sethy, and B. Ramabhadran,“Balancing false alarms and hits in spoken term detection,” in *Proc. ICASSP*, 2010.

[7] I. Bulyko, O. Kimball, M.H. Siu, J. Herrero, and D. Blum, "Detection of unseen words in conversational mandarin," in *Proc. ICASSP*, 2012

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

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

[10] J. Tejedor, M. Fapso, I. Szoke, J. Cemocky and F. Grezl, “Comparison of methods for language-dependent and language-independent query-by-example spoken term detection,” *ACM Transactions on Information Systems*, 2012.

[11] C. Chan and L.S. Lee, “Model-based unsupervised spoken term detection with spoken queries,” *IEEE Transactions on Audio Speech & Language Processing*, pp. 1330-1342, 2013.

[12] H. Wang, T. Lee, CC. Leung, B. Ma and H. Li, “Acoustic segment modeling with spectral clustering methods,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, pp. 264–277, 2015.

[13] I. Szoke, L. Burget, F. Grezl, JH. Cernocky, L. Ondel, “Calibration and fusion of query-by-example systems — But SWS 2013,” In *Proc. ICASSP*, pp. 7849 – 7853, 2014.

[14] X. Anguera, et al. “Query-by-example spoken term detection on multilingual unconstrained speech,” In *Proc. Interspeech*, pp. 2459–2563, 2014.

[15] H. Xu, P. Yang, X. Xiao, L. Xie, “Language independent query by example spoken term detection using N-best phone sequence and partial matching,” In *Proc. ICASSP*, pp. 5191 – 5195, 2015.

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

[17] G. Chen, C. Parada and T.N. Sainath, “Query by example keyword spotting using long short-term memory network,” in *Proc. ICASSP*, 2015.

[18] H. Wang, T. Lee, and C.C. Leung, "Unsupervised spoken term detection with acoustic segment model," in *Proc. Speech Database and Assessments (Oriental COCOSDA)*, pp. 106 – 111, 2011.

[19] Y. Qiao, N. Shimomura, and N. Minematsu. "Unsupervised optimal phoneme segmentation: Objectives, algorithm and comparisons," in *Proc. ICASSP*, pp. 3989-3992, 1988.

[20] X. Anguera, L.J. Rodriguez-Fuentes, A. Buzo, F. Metze, I. Szoke and M. Penagarikano. "QUESST2014: Evaluating query-by-example speech search in a zero-resource setting with real-life queries," in *Proc. ICASSP*, pp. 5833-5837, 2015.

[21] I. Szoke, L.J. Rodriguez-Fuentes, A. Buzo, X. Anguera, F. Metze, J. Proenca, M. Lojka, and X. Xiong, “Query by example search on speech at MediaEval 2015," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, Wurzen, Germany, 2015.

[22] M. Muller, *Information retrieval for music and motion*, Springer-Verlag, 2007.