Interspeech 2018
2-6 September 2018, Hyderabad

Siamese Recurrent Auto-Encoder Representation for Query-by-Example Spoken Term Detection

Ziwei Zhu[1], *Zhiyong Wu*[1,2], *Runnan Li*[1], *Helen Meng*[1,2], *Lianhong Cai*[1]

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, China
[2]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, China
{zhuzw15, lirn15}@mails.tsinghua.edu.cn,
{zywu, hmmeng}@se.cuhk.edu.hk,clh-dcs@tsinghua.edu.cn

Abstract

With the explosive development of human-computer speech interaction, spoken term detection is widely required and has attracted increasing interest. In this paper, we propose a weak supervised approach using Siamese recurrent auto-encoder (RAE) to represent speech segments for query-by-example spoken term detection (QbyE-STD). The proposed approach exploits word pairs that contain different instances of the same/different word content as input to train the Siamese RAE. The encoder last hidden state vector of Siamese RAE is used as the feature for QbyE-STD, which is a fixed dimensional embedding feature containing mostly semantic content related information. The advantages of the proposed approach are: 1) extracting more compact feature with fixed dimension while keeping the semantic information for STD; 2) the extracted feature can describe the sequential phonetic structure of similar sounds to degree, which can be applied for zero-resource QbyE-STD. Evaluations on real scene Chinese speech interaction data and TIMIT confirm the effectiveness and efficiency of the proposed approach compared to the conventional ones.

Index Terms: Siamese recurrent auto-encoder, query-by-example spoken term detection, zero-resource

1. Introduction

With the explosive development of human-computer speech interaction, spoken term detection tasks have drawn increasing interests, e.g. detecting specific segments from massive speech utterances in search engine, wake-up and command control in smart home hardware, smart phones and vehicular network. Spoken term detection tasks can be categorized into two kinds according to the terms format: 1) keyword search using text; 2) query-by-example spoken term detection (QbyE-STD) using speech segments. Unlike keyword search [1][2] which is mainly based on large vocabulary continuous speech recognition (LVCSR), QbyE-STD system consists of two parts: feature extraction and term detection. The feature extraction part among them directly affects the overall performance of the QbyE-STD system, because speech contains rich and variable characteristics such as semantic content, speaker characteristics, environment noise, etc. In this paper, we explore a compact semantic content related feature representation for QbyE-STD.

Different from the semantic meaning in the text, the semantic content derived from speech signals mentioned in the paper is referred to the sequence of phonetic structure. We usually use the phonemes sequence to describe the speech content in one language, but many phonemes are not shared between different languages. Similar sounds between different languages can be modeled by smaller modeling units than phonemes of international phonetic alphabet (IPA). These modeling units are unknown to us and usually represented by the neural networks (NN). Multi-language bottleneck features [3] [4] have been used to represent the shared cross-language information by discriminative frame-level features for QbyE-STD. However, we think the NN modeling of entire audio sequence can ignore the frame-level mismatches of pronunciation representation between different languages, so we wish to use sequential phonetic structure to represent the semantic content derived from speech segments.

It has always been a research focus that extracts compact features for variable length vectors sequence. Recurrent neural networks (RNN) with long short term memory (LSTM) can represent the input segment by using the outputs of the hidden layer of RNN at the last few time steps for QbyE-STD, which is trained by targeting to word units [5]. However, the acquired feature only models the distributed acoustic word embedding space rather than the sequential phonetic structure. In addition, one of the vital problems in these supervised methods is the requirement of a large amount of labeled data that is hard to collect. Recurrent auto-encoder (RAE) [6], which is called sequence-to-sequence auto-encoder (SA) in [7], has been successful for mapping variable-length audio segments into fixed dimensional vectors in an unsupervised way. But the vector representation still contains non-semantic information such as speaker and emotion, which maybe influence the performance of the QbyE-STD system. Hence, weak supervised methods using pairwise information have attracted increasing interest of researchers. [8] proposed the use of Siamese convolutional neural network (CNN) to learn shared semantic representation of one word from various speakers while increasing the representation differences between different words. However, the approach does not possess the strong ability in capturing long term temporal dependences for modelling the sequential phonetic structure.

In this paper, we proposed a weak supervised method using Siamese RAE to represent variable-length audio segments by fixed dimensional vectors that are mostly related to semantic content for QbyE-STD. Different from the previous work in RAE [6][7], pairwise information helps learning Siamese RAE, which is inspired from the research on semantic similarity[9]. We hope the vector representations obtained in this method can describe more precisely the sequence of phonetic structure, so the representation learned by high-resource language is also

10.21437/Interspeech.2018-1788

effective in zero-resource QbyE-STD. This is referred to Siamese RAE representation for QbyE-STD. QbyE-STD using Siamese RAE here only needs the similarities between two single vectors, which is much more efficient than the conventional dynamic time warping (DTW) based approaches. Experimental results demonstrate the superior performance of the proposed approach.

The rest of the paper is organized as follows. Section 2 describes in detail the proposed Siamese RAE model. Section 3 introduces the Siamese RAE based QbyE-STD system. Experiments and analyses are provided in the Section 4. Section 5 concludes the paper.

2. Siamese RAE Representation

The proposed Siamese RAE model is outlined in Figure 1. There are two networks RAE_a and RAE_b which each processes one of the segments of a given pair, and the Siamese architecture is the two networks with tied parameters, which means RAEa = RAEb in this work. It's desired that the encoder of the RAE learns a mapping from the space of variable length acoustic feature vectors $X = \{x_1, x_2, \dots, x_T\}$ into a fixed dimensional vector z, which can describe the semantic content (the sequential phonetic structure) of the segment to the maximum extent. The learned vector representation is called Siamese RAE representation. In addition, the computational time of the subsequent detection process in QbyE-STD can be reduced at large due to the single vector representation.

2.1. RAE model

We use the RAE model in [6], which consists of three parts as shown in Figure 1: 1) An encoder RNN with gated recurrent unit (GRU) hidden units; 2) A dense layer with weight matrix W and D rectified linear units (ReLUs) denoted by g; 3) A decoder GRU-RNN. The encoder RNN reads the input acoustic feature vectors sequence $X = \{x_1, x_2, \dots, x_T\}$ sequentially and the hidden state of the RNN is updated accordingly. The dense layer g then transforms the hidden state vector h_T^{ε} from time step T into a D-dimensional vector $z = g(Wh_T^{\varepsilon})$. The decoder GRU-RNN takes the vector z as input at each time step t to reconstruct the original sequence of T acoustic feature vectors. The D-dimensional vector z is extracted as the feature representation for the input speech segment.

2.2. Semantic content similarity for Siamese RAE

To make the extracted feature z more related to the semantic content, we apply a pre-defined similarity function to the vector representations of the segments of the given pair. Motivated by the previous work [8], we use a margin-based (hinge) loss which is based on the computation of the relative distance, in which the intra class distance can be smaller while the distance between classes become larger. The objective function of Siamese RAE is a weighted sum of the reconstruction loss for each segment and the similarity loss between the vector representations for segments. In this way, the Siamese RAE is able to make the vector representation z learn the semantic content as much as possible while depressing other non-semantic information.

The features extracted from each segment $\mathrm{X}_i = \{x_{i1}, x_{i2}, \dots, x_{iT}\}$, where $x_{it} \in \mathbb{R}^{\mathrm{d}}, d = 40$ are sent to the RAE and output $\mathrm{X}_{\mathrm{o}} = \{x_{o1}, x_{o2}, \dots, x_{oT}\}$ respectively. For each group of inputs, there are two pairs composed of three feature

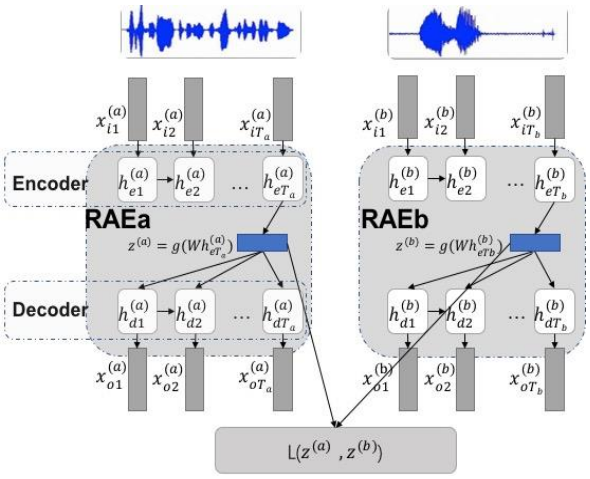

Figure 1: *Siamese recurrent auto-encoder (RAE) architecture in which RAEa (the left large block) and RAEb (the right large block) are with tied parameters. The encoder of RAE maps the speech segment (variable length input feature vectors sequence) into a fixed dimensional vector representation z (the blue box). The whole model is trained by the weighted sum of the segments' reconstruction losses and the similarity loss between the vector representations for these segments.*

vectors sequence: anchor sample $\mathrm{A}_i = \{x_{i1}^{(a)}, x_{i2}^{(a)}, \dots, x_{iTa}^{(a)}\}$, positive sample $\mathrm{P}_i = \{x_{i1}^{(p)}, x_{i2}^{(p)}, \dots, x_{iTp}^{(p)}\}$ (of the same semantic content type with the anchor sample) and negative sample $\mathrm{N}_i = \{x_{i1}^{(n)}, x_{i2}^{(n)}, \dots, x_{iTn}^{(n)}\}$ (of the different semantic content type with the anchor sample). The reconstruction losses of segments A, P and N can then be computed as the mean square error (MSE) between the original features and the recovered features.

$$MSE_{loss}A = \sum_{k=1}^{Ta} \frac{(x_{ik}^{(a)} - x_{ok}^{(a)})^2}{Ta \times d} \quad (1)$$

$$MSE_{loss}P = \sum_{k=1}^{Tp} \frac{(x_{ik}^{(p)} - x_{ok}^{(p)})^2}{Tp \times d} \quad (2)$$

$$MSE_{loss}N = \sum_{k=1}^{Tn} \frac{(x_{ik}^{(a)} - x_{ok}^{(a)})^2}{Tn \times d} \quad (3)$$

The hinge loss is used to measure the relative semantic similarity between the segments' vector representations $z^{(a)}$, $z^{(p)}$ and $z^{(n)}$.

$$Hinge_{loss} = \max\{0, M + l(z^{(a)}, z^{(p)}) - l(z^{(a)}, z^{(n)})\} \quad (4)$$

where the similarity distance l between two vector representations z_1 and z_2 is computed by the cosine distance as follows:

$$l(z_1, z_2) = (1 - cos(z_1, z_2))/2) \quad (5)$$

Finally, the objective function of Siamese RAE is the weighted sum of the three segments' recovery losses and the semantic hinge loss as follows and will be back propagated through the whole model:

$$total_{\text{loss}} = (1-a)Hinge_{loss} + a(MSE_{loss}A + MSE_{loss}P + MSE_{loss}N)/3 \quad (6)$$

where a is the weight used to balance the semantic similarity loss and the recovery losses.

Note that the vector representation z trained by Siamese RAE is the Siamese RAE representation used for QbyE-STD.

3. Siamese RAE based QbyE-STD

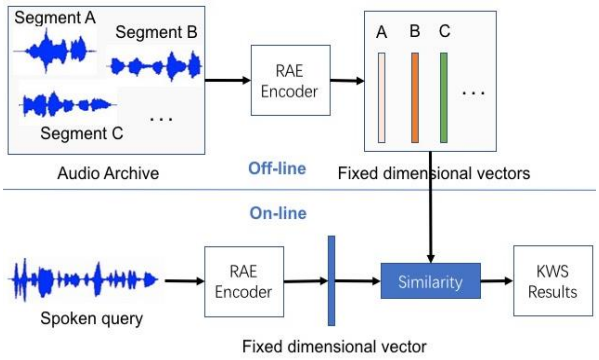

Figure 2: *Siamese RAE based QbyE-STD framework*

The audio segment representation learned by the proposed Siamese RAE model can be applied to QbyE-STD as shown in Figure 2, which is inspired from the previous work [7]. The system consists of off-line and on-line processes. The off-line process is in the upper half of Figure 2. First, the audio archives are segmented based on word boundaries into variable-length speech segments. As for the speech of zero-resource language, voice activity detection (VAD) is available for segmenting. Then, the system exploits the trained RAE encoder in Figure 1 to encode these variable length segments into fixed dimensional vectors. When a spoken query is entered in the lower left corner of Figure 2, the same RAE encoder can encode the input spoken query similarly into a fixed dimensional vector. The evaluation of QbyE-STD is based on the cosine similarities computed and ranked between the vector representation of the query and those of all segments in the archive. Please note that the computation consumption of the on-line process here is very low, due to the similarity comparison between single vector representations.

4. Experiments

4.1. Experiments Setup

We use a real scene Chinese speech interaction data from Sogou and TIMIT for the experiments. 40 dimensional F-bank features with mean and variance normalization (MVN) on each speech segment are extracted using Kaldi toolkit [10] and used as the original acoustic features.

- There are totally 500,000 speech sentences randomly selected from Sogou human-computer speech interaction engine for the experiments. All 500,000 sentences have been force aligned to get the duration for each segment that contains a complete word meaning. We select segments which consist of at least 6 phonemes and are with the length between 0.5 and 2.0 seconds. There are 50K segments in the training set, which belong to 18,320 different word types. For each segment, we select randomly a segment of the same word type in the training set to form a positive pair, and a segment of different word type to form a negative pair. Ultimately, there are 50K groups (100K pairs) in total for training.
- In the Chinese test set, there are totally 24,916 word types, of which 6,596 word types (about one-quarter) are not in the training set, so we regard them as OOVs. We randomly select 3,723 segments as the Chinese speech archive, in which one-third are OOVs, and 3,723 segments as spoken queries for QbyE-STD. The word type of each query only occurs once in the archive.
- TIMIT is regarded as the zero-resource language dataset and then used to evaluate the performance of the learned representation in a different language. For comparison, the segmentation is based on word boundaries instead of VAD. Thereafter, the speech segments are grouped into different word types, from which the functional words such as articles, conjunctions are removed. Finally, 2,089 word types that occur most frequently in the TIMIT dataset are selected, from which two speech segments are randomly selected for each word type, one to form the English speech archive and the other as the spoken query.

Deep learning models in this paper are implemented with Pytorch [11]. Mini-batch-trained Adam [12] with 0.0001 learning rate and 40 batch size is used for training. All RNN based models are with GRU units and trained for 2 epochs. The number of the RAE encoder hidden layer units and D in section 2.1 are set to 300 as the previous work in [6]. Mean Average Precision (MAP), which is the mean of the average precision in the range of recall for each query in the test set, was used as the evaluation metrics for QbyE-STD as the previous work [3]. Several approaches implemented for the experiments are described as follows:

- F-bank (DTW-based): the mostly used baseline with the help of frame-based DTW [13] for spoken term detection, the F-bank features are processed with MVN.
- Siamese CNN: the best of the previous approaches on word discrimination task [8], which obtained acoustic word embeddings from padded speech input by CNN. All hyper-parameters in the approach are set as the previous work [8].
- GRU-RAE: an unsupervised method with GRU as hidden units to represent variable length vectors sequence by a fixed-length vector, which is equivalent to the model by setting a of Eq.6 to 1 in section 2.2.
- Siamese RNN: The model by setting $a = 0$ for Eq.6. By using pairwise information, the model exploits LSTM to map variable length vectors sequence into a fixed dimensional embedding space.
- Siamese RAE: The proposed model, which combines the advantages of GRU-RAE and Siamese RNN. The a of Eq.6 is set to 0.5.

4.2. Analysis of the sequential phonetic structure learned by Siamese RAE in Chinese

Table 1 shows the average cosine distance of segment pairs with different phoneme sequence edit distance and suffixes for analysis of the sequential phoneme structure. These pairs are randomly selected from the Chinese test set. Due to the vector representation is extracted from the last hidden state of the RNN, we distinguish pairs with the same or different suffixes (the last phoneme). For the segment pairs with the same suffixes, there are 49,868 pairs whose phoneme sequence edit distances are less than 5, and 65,378 pairs whose edit distances are equal to 5. While for the pairs with different suffix, 57,836 pairs are with edit distance less than 5, and 106,735 pairs with edit distance equal to 5. From Table 1, it can be seen, for the learned vector representation of all three RNN derived networks, the average

cosine distance of segment pairs is increasing with the phoneme sequence edit distance increased, which implies the RNN-based model can describe the sequential phonetic structure to some extent. An interesting observation is that the cosine distance of the learned representation for the pairs with same suffixes is less than that of pairs with different suffixes. This demonstrates that the last phoneme of the word is indeed important in the sequential phonetic structure. Another important observation is that, compared to Siamese RNN, the proposed Siamese RAE model can reduce the gap of cosine similarity distances between the segment pairs with the same and different suffixes, due to the reconstruction mechanism of RAE.

Table 1: *Average cosine distance of the learned representation between segment pairs clustered by the phoneme sequence edit distance, on the condition of whether the suffixes (the last phoneme) are the same.*

Phoneme Sequence Edit Distance	Average cosine distance of the learned representation for segment pairs with same / different suffixes (difference in brackets)		
	GRU-RAE	**Siamese RNN**	**Siamese RAE**
<5	0.0321 / 0.0326 (+0.004)	0.769 / 0.724 (-0.045)	0.694 / 0.653 (-0.041)
5	0.0326 / 0.0333 (+0.007)	0.926 / 0.848 (-0.078)	0.826 / 0.752 (-0.074)

4.3. Evaluation of QbyE-STD on the language that is the same as the one the representation is learned from

Table 2: *Performance of Siamese RAE and its comparisons on QbyE-STD task (on the same language the representation is learned).*

Model	Dim	MAP (IVs)	MAP (OOVs)	MAP (total)
F-bank (DTW-based)	40	0.048	0.057	0.051
Siamese CNN	1024	0.083	0.091	0.086
GRU-RAE	300	0.060	0.065	0.061
Siamese RNN	300	0.082	0.084	0.083
Siamese RAE	300	**0.112**	**0.126**	**0.116**

We divide the test set for QbyE-STD into In-Vocabulary (IVs) test set and Out-Of-Vocabulary (OOVs) test set in accordance with the belonging of the entered spoken query in the training set. Table 2 shows the performance of Siamese RAE and all comparisons on the Chinese testing set. The F-bank (DTW-based) model is based on Frame-based DTW approach. We can see from the table that the performance of the Siamese CNN is better than the Siamese RNN on the STD task. The possible reason may be the dimension of the learned representation is 1024 for Siamese CNN, containing more useful information for STD. GRU-RAE, a totally unsupervised method, is better than F-bank method as it can learn more representative information by auto-encoders, while is worse than Siamese RNN because of the missing of pairwise information. As expected, Siamese RAE performs the best among all the approaches, indicating that it can generate vector representations containing both representative information presented by the reconstruction function of RAE model and distinctive information introduced by the Siamese architecture at the same time.

4.4. Evaluation of QbyE-STD on the language that is different from the one the representation is learned

In the above experiment, for all approaches, the performance on OOVs test set is similar to the IVs. The right judgments on OOV segments indicate that the model can learn the sequential phonetic structure for similar sounds. Inspired by the origin of the IPA, we assume the vector representation may be also effective in other languages, although trained on Chinese.

Table 3: *Performance of Siamese RAE and its comparisons on QbyE-STD task (on a language different from the one where the representation is learned).*

Model	MAP (2 epochs)	MAP (5 epochs)
F-bank (DTW-based)	0.241	0.241
GRU-RAE	0.242	0.333
Siamese RNN	0.178	0.065
Siamese RAE	**0.242**	**0.234**

To verify the idea, we test the vector representation learned from Chinese on an English dataset TIMIT. Table 3 shows the performance of the proposed Siamese RAE and comparisons on the TIMIT dataset. It's obvious that the performance of GRU-RAE is better than the DTW-based approach while Siamese RNN is worse, which attests that pairwise information in Chinese is detrimental to the model directly used for English spoken term detection. However, the proposed Siamese RAE model, which combines the advantages of Siamese RNN and GRU-RAE, behaves between the GRU-RAE and Siamese RNN. Due to the similar sound between Chinese and English is limited, the vector representation learned by GRU-RAE performs better than Siamese RAE with the increase of the training epochs. It indicates the sequential phonetic structure can be learned by Siamese RAE and the performance of the model depends on the similarity of the languages. In conclusion, not only the Siamese RAE representation performs best on the same language dataset, but also is effective in different languages, which can be applied for zero-resource QbyE-STD.

5. Conclusion

In this paper, we propose a Siamese adaptation of RAE model for QbyE-STD. By using Siamese RAE to map variable length speech segments into fixed dimensional vectors, the learned feature representation can describe the semantic content (the sequential phonetic structure) to some extent. Furthermore, due to similar sounds between different languages, the sequential phonetic structure learned by one language is also effective in another different languages, which can be used for zero-resource QbyE-STD. In addition, the detection time can be reduced at large due to the single vector representation. Evaluations on real scene Chinese speech interaction data and TIMIT confirmed the effectiveness and efficiency of the proposed approach in spoken term detection task.

Although Siamese RAE performs well in the task, the encoder model is too simple to describe the sequential phonetic structure at large. In the future, we will investigate an improved version of the proposed approach.

6. Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC) (61433018), joint research fund of NSFC-RGC (Research Grant Council of Hong Kong) (61531166002, N_CUHK404/15) and National Social Science Foundation of China (13&ZD189). The work was conducted when the first author was an intern student visiting Beijing Sogou Science and Technology Development Co., Ltd.

7. References

[1] B. Gundogdu, M. Saraclar, “Similarity learning based query modeling for keyword search,” in *Proc. Interspeech*, pp. 3617–3621, 2017.

[2] Z. Chen, J. Wu, “A rescoring approach for keyword search using lattice context information,” in *Proc. Interspeech*, pp. 3592-3596, 2017.

[3] H. Chen, Cheung-Chi Leung, L. Xie, B. Ma, and H. Li, “Unsupervised bottleneck features for low-resource query-by-example spoken term detection,” in *Proc. Interspeech*, pp. 923-937, 2016,.

[4] Y. Yuan, Cheung-Chi. Leung, L. Xie, H. Chen, B. Ma, H. Li, “pairwise learning using multi-language bottleneck feature for low-resource query-by-example spoken term detection”, in *Acoustics, Speech and Signal Processing, 2017 IEEE International Conference on*, 2017.

[5] G. Chen, C. Parada, T. N. Sainath, "Query-by-example keyword spotting using long short-term memory networks", in *Acoustics, Speech and Signal Processing, 2015 IEEE International Conference on*, pp. 5236-5240, 2015.

[6] E. Audhkhasi, A. Rosenberg, A. Sethy, B. Ramabhadran, B.Kingsbury, “End-to-end ASR-free keyword search from speech,” in *IEEE Journal of Seclected Topics in Signal Processing*, pp.1351-1359., 2017.

[7] Y. Chung, C. Wu, C. Shen, H. Lee, L. Lee, “Audio word2vec: unsupervised learning of audio segment representations using sequence-to-sequence autoencoder”, in *Proc. Interspeech*, 2016.

[8] H. Kamper, W. Wang, K. Livescu, "Deep convolutional acoustic word embeddings using word-pair side information", in *Acoustics, Speech and Signal Processing, 2016 IEEE International Conference on*, pp. 4950-4954, 2016.

[9] Jonas Mueller and Aditya Thyagarajan. “Siamese recurrent architectures for learning sentence similarity”. In *Proc. AAAI.*, pp. 2786–2791, 2016.

[10] Povey, et al, “The Kaldi Speech Recognition Toolkit”, in *Proc. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[11] Pytorch. [OL]. [2017-10-19]. http://pytorch.org/docs

[12] D. Kingma, J. Ba, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2014.

[13] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 1, pp. 43–49, 1978.