Interspeech 2018
2-6 September 2018, Hyderabad

Learning Acoustic Word Embeddings with Temporal Context for Query-by-Example Speech Search

Yougen Yuan[1], *Cheung-Chi Leung*[2], *Lei Xie*[1*], *Hongjie Chen*[1], *Bin Ma*[2], *Haizhou Li*[3]

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Alibaba Inc., Singapore
[3]Department of Electrical and Computer Engineering, National University of Singapore, Singapore

{ygyuan,hjchen}@nwpu-aslp.org, lxie@nwpu.edu.cn,
{cc.leung,b.ma}@alibaba-inc.com, haizhou.li@nus.edu.sg

Abstract

We propose to learn acoustic word embeddings with temporal context for query-by-example (QbE) speech search. The temporal context includes the leading and trailing word sequences of a word. We assume that there exist spoken word pairs in the training database. We pad the word pairs with their original temporal context to form fixed-length speech segment pairs. We obtain the acoustic word embeddings through a deep convolutional neural network (CNN) which is trained on the speech segment pairs with a triplet loss. By shifting a fixed-length analysis window through the search content, we obtain a running sequence of embeddings. In this way, searching for the spoken query is equivalent to the matching of acoustic word embeddings. The experiments show that our proposed acoustic word embeddings learned with temporal context are effective in QbE speech search. They outperform the state-of-the-art frame-level feature representations and reduce run-time computation since no dynamic time warping is required in QbE speech search. We also find that it is important to have sufficient speech segment pairs to train the deep CNN for effective acoustic word embeddings.

Index Terms: acoustic word embeddings, word pairs, temporal context, triplet loss, query-by-example spoken term detection

1. Introduction

Query-by-example (QbE) speech search or spoken term detection is the task of searching for the occurrence of a spoken query in search content [1, 2]. A typical approach to this task relies on dynamic time warping (DTW) to perform acoustic pattern matching over frame-level feature representations. These feature representations can be learned in unsupervised [3, 4, 5, 6] or supervised [7, 8, 9] manner. In supervised learning, classifiers are usually trained using labeled data from non-target languages to derive features.

In this paper, we propose to learn acoustic word embeddings with temporal context for QbE speech search. We assume that there exist spoken word pairs in the training database in the target language. Learning frame-level feature representations using word pairs has been shown successful [10]. However, it relies on computationally expensive DTW during the search. This prompts us to study acoustic word embeddings that encode speech at segment or word level. In this way, we can simplify the QbE speech search test by measuring the vector distance (e. g., cosine distance) over the acoustic word embeddings between the spoken query and the search content.

Acoustic word embeddings project speech segments into fixed-dimensional vector space where the distance between same speech content is small while the distance between different speech content is large. They have been shown successful in automatic speech recognition [11] and isolated word discrimination [12, 13, 14]. However, as the word boundary is not available in search content, QbE speech search is therefore considered more difficult than isolated word discrimination. To overcome this problem, studies have shown [15, 16] that approximate nearest neighbor search over the acoustic word embeddings is a reasonable solution.

We propose the idea to include the leading and trailing word sequences as the temporal context of a word. The word pairs are padded with their original temporal context to form fixed-length speech segment pairs. We train a deep convolutional neural network (CNN) with a triplet loss using the speech segment pairs to learn acoustic word embeddings. During QbE speech search, we propose to shift a fixed-length analysis window to obtain a sequence of embeddings on search content, then we search over the embeddings instead of frame-level feature representations to find the matching spoken query.

The novel contribution of this paper is that, for the first time, we incorporate the temporal context to improve the acoustic word embeddings for QbE speech search. With the temporal context, we learn the possible neighboring speech sequences around the words, which reduces the mismatch between the learning of embeddings and its application on the search content. Our proposed technique outperforms the state-of-the-art frame-level feature representations [10] and reduces the run-time computation since no DTW is required in QbE speech search. As learning acoustic word embeddings requires a larger number of word pairs than learning frame-level feature representations [17], we use more word pairs discovered from the Switchboard speech corpus and we test the effect of number of speech segment pairs in learning acoustic word embeddings for QbE speech search.

2. QbE speech search using acoustic word embeddings

2.1. Learning embeddings with temporal context

Acoustic word embeddings which are extracted from a typical feed-forward deep neural network usually require

* Corresponding author

10.21437/Interspeech.2018-1010

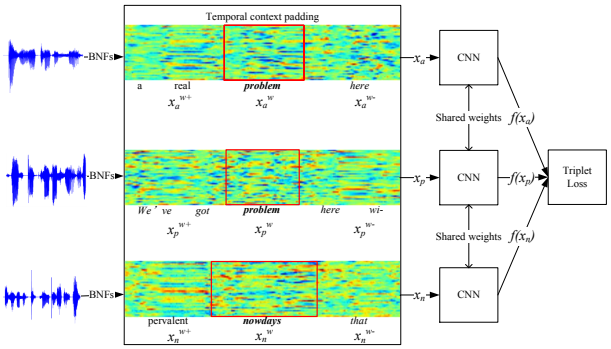

Figure 1: *The diagram of learning acoustic word embeddings with temporal context padding.*

the input with fixed-length. Zero padding has been shown successful in learning acoustic word embeddings for isolated word discrimination [13, 17]. With zero padding, all speech segments are padded with zeros on both side of each segment to the same length. In the case of QbE speech search, because the word boundary is not available, it is hard to segment the search content into isolated words. Therefore, we propose to shift a fixed-length analysis window to segment the search content into many fixed-length speech segments. As these speech segments may contain sub-words, one or more whole words, there exists a mismatch between the learning of embeddings and the use of embeddings at run-time. To mitigate such mismatch, we propose to use the temporal context of a word to learn acoustic word embeddings in QbE speech search.

The temporal context refers to the original leading and trailing word sequences on both sides of a word. The way we incorporate the temporal context is also called temporal context padding. As illustrated in Fig. 1, given a word instance ${x_a}^w$, we add its original previous word sequence as ${x_a}^{w+}$ in front and its subsequent word sequence as ${x_a}^{w-}$ behind with the same number of frames. Notice that the temporal context may contain a partial word (e. g., "wi-" in "with"), a whole word (e. g., "here"), or multiple words (e. g., "a real"). We assume that word pairs (e. g., $({x_a}^w, {x_p}^w)$) identified by humans are available. The word pairs are padded with their original temporal context to form speech segment pairs (e. g., (x_a, x_p)). In this way, the speech segment x_a contains the same word (e. g., "problem") as the speech segment x_p, while the speech segment x_n contains a different word (e. g., "nowadays").

We employ deep neural networks with a triplet loss to learn acoustic word embeddings. The deep neural networks take the triplets as input. Each triplet consists of 3 examples (x_p, x_a, x_n). We use a pair of speech segments as an anchor example x_a and a positive example x_p, and we randomly sample another speech segment as a negative example x_n. Learning acoustic word embeddings is shown in Fig 1. Each example is represented by multi-lingual bottleneck features (BNFs), which capture rich information of phonetic discrimination from other language resources. We aim to learn a function f which maps an example x to the fixed-dimensional embedding $f(x)$. The deep CNN, which has been shown successful to learn this function in isolated word discrimination [13, 17], is used here for learning embeddings with temporal context.

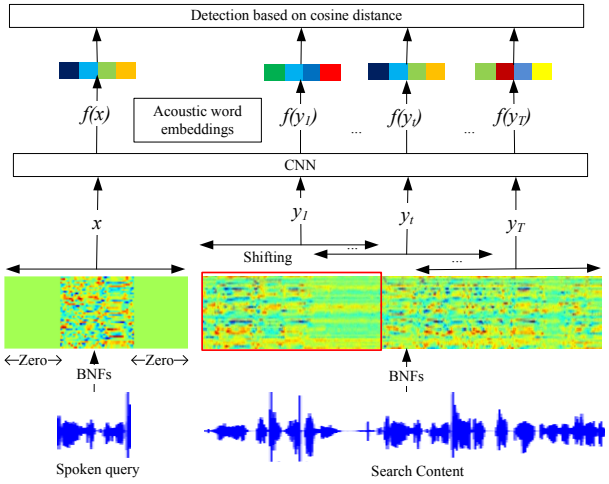

Figure 2: *QbE speech search using acoustic word embeddings.*

The triplet loss was originally proposed in [18] for learning face embeddings from image. As for learning acoustic word embeddings from speech, we aim to increase the similarity between the embeddings $(f(x_p), f(x_a))$, while decreasing the similarity between the embeddings $(f(x_n), f(x_a))$. Our triplet loss is defined as

$$Loss(x_p, x_a, x_n) = max\{0, \delta + d^+ - d^-\} \quad (1)$$

$$d^+ = \frac{1 - \frac{f(x_p)\cdot f(x_a)}{\|f(x_p)\|_2 \|f(x_a)\|_2}}{2} \quad (2)$$

$$d^- = \frac{1 - \frac{f(x_n)\cdot f(x_a)}{\|f(x_n)\|_2 \|f(x_a)\|_2}}{2} \quad (3)$$

where δ is a margin constraint that regularizes the gap between the cosine distance of same-word embeddings d^+ and the cosine distance of different-word embeddings d^-. We set the margin to 0.15 as in [13, 17]. The acoustic word embeddings $f(x)$ are extracted from the last layer of the trained deep CNN.

2.2. Shifting analysis window on search context

Fig. 2 illustrates the process of our proposed QbE speech search system using acoustic word embeddings. As an indexing process, we propose to apply a fixed-length analysis window on the search content y by shifting along the time axis. The speech segment in the analysis window is then converted into an acoustic word embedding by the trained deep CNN. As a result, the search content is indexed by a sequence of acoustic word embeddings as $(f(y_1), ..., f(y_i), ..., f(y_T))$. As no context information is available for the spoken query x, we pad zeros to both sides of x as the input to the trained deep CNN to obtain the embedding $f(x)$. In this way, the vector distance, instead of DTW distance, can be directly used over the embeddings between the spoken query and the search content.

Note that the mismatch between the embeddings in the spoken query and the search content still exist. To further mitigate the mismatch, we also considered different ways to learn the embeddings, including zero padding, and combining context padding and zero padding (using the two padding methods in a speech segment triplet or using different padding

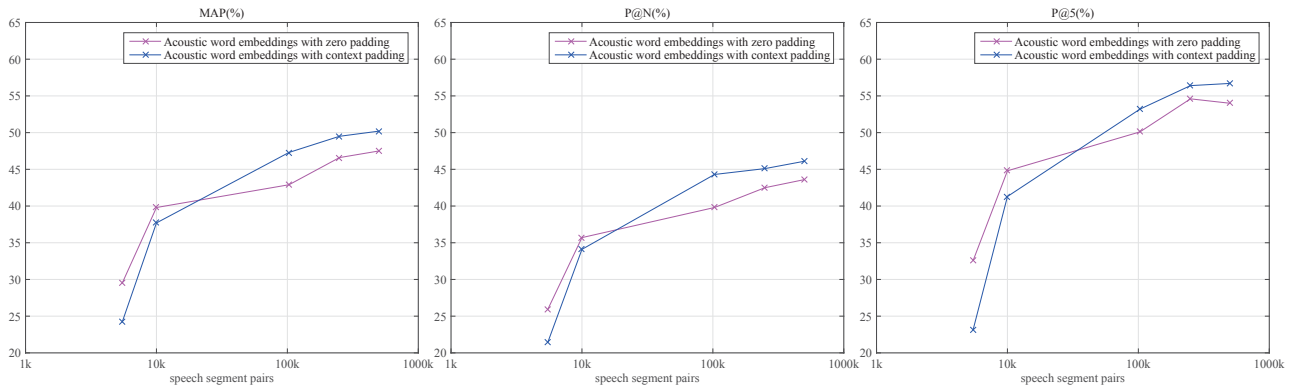

Figure 3: *Comparative study between zero padding and context padding in acoustic word embeddings for QbE speech search. Multilingual BNFs and Set 2 are used.*

methods in different triplets). However these ways do not further improve the search accuracy in our preliminary test.

The size of the fixed-length analysis window is determined by the average length of all speech segments used in learning acoustic word embeddings. We set the window shift size to 5 frames as we find that a shifting smaller than 5 frames doesn't improve. A minimum distance cost can be calculated by:

$$Cost(x,y) = min(1 - \frac{f(x) \cdot f(y_i)}{\|f(x)\|_2 \|f(y_i)\|_2}), i = 1, ..., T \quad (4)$$

Given a spoken query, all the minimum distance costs in search content are returned by the QbE speech search system.

3. Experiments

3.1. Experimental setup

To evaluate the effectiveness of our proposed acoustic word embeddings, we conducted the QbE speech search on the English Switchboard corpus. From our previous work [17], we observed that learning word-level embeddings should require a larger number of word pairs than learning frame-level feature representations. Thus we extended two training sets:

- Set 1: It has the same vocabulary size (1,687) as in [17], but it consists of $37k$ word instances (about 6.6 hours of speech) that make up $500k$ speech segment pairs with temporal context padding.
- Set 2: The vocabulary size is increased to 5,476, and the dataset consists of $53k$ word instances (about 9.5 hours of speech) that also make up $500k$ speech segment pairs with temporal context padding.

We used the same development set as in [19, 20, 13, 17, 10] for learning acoustic word embeddings. As for QbE speech search tests, we followed the data setting in [10]. We used 346 spoken queries as the keyword set and 100 utterances as the test set.

We included the leading and trailing word sequences as the temporal context of a word to form a speech segment with the length of 200 frames (2 seconds). All the speech segments were represented by 40-dimensional multi-lingual BNFs as in [10]. The BNF extractor was trained using Mandarin Chinese and Spanish telephone speech. We trained a deep CNN with a triplet loss using speech segment pairs to learn acoustic word embeddings. The deep CNN model consists of two convolutional and max pooling layers, a fully-connected ReLU layer with 2,048 hidden units and a fully-connected linear layer with 1,024 hidden units. We implemented the model using the Theano toolkit [21], and we trained the model using stochastic gradient descent with the mini-batch size of 1,024. All the neural weights were initialized randomly. An ADADELTA [22] optimizer was used with the momentum of $\rho = 0.9$ and the precision of $\epsilon = 10^{-6}$. Training would be terminated if the loss on the development set was not improved over 20 epochs.

As in [3, 6, 10], three different evaluation metrics are used for QbE speech search: 1) mean average precision (MAP), which is the mean of average precision for each query on search content. 2) Precision of the top N utterances in the test set (P@N), where N is the number of target utterances involving the query term. 3) Precision of the top 5 utterances in the test set (P@5).

3.2. Temporal context in acoustic word embeddings

To validate the efficiency of our proposed acoustic word embeddings learned with temporal context for QbE speech search, we compared context padding with zero padding in learning acoustic word embeddings based on multi-lingual BNFs. From Fig. 3 we can find that context padding outperforms zero padding when more than $10k$ speech segment pairs are available in Set 2 (about 3%-11% relative improvement in three evaluation metrics). Similar results are also obtained in Set 1 with a smaller vocabulary. The experiment results suggest that the temporal context padding learns the possible neighboring speech sequences around the words, which reduces the mismatch between the learning of embeddings and the use of embeddings on the search content for QbE speech search.

3.3. Comparison of different feature representations

Table 1 lists the performance of QbE speech search using different feature representations, including multi-lingual BNFs, the state-of-the-art frame-level feature representations in [10] and our proposed acoustic word embeddings learned with temporal context. These feature representations are trained using $500k$ speech segment pairs in Set 2. The QbE speech

Table 1: *Comparison of different feature representations for QbE speech search. 500k speech segment pairs in Set 2 are used.*

Representation	Input features of paired examples	Use DTW?	Run-time computation (seconds)	MAP	P@N	P@5
Multi-lingual BNFs	N/A	Yes	4,752	0.400	0.365	0.485
Frame-level feature representations [10]	Multi-lingual BNFs	Yes	9,506	0.485	0.446	0.566
Acoustic word embeddings (proposed)	Multi-lingual BNFs	No	**1,017**	**0.502**	**0.462**	**0.567**

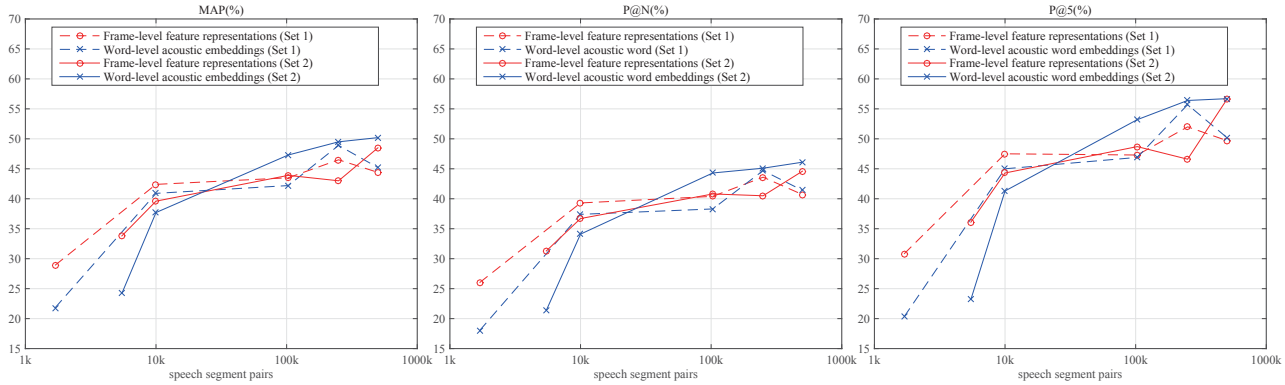

Figure 4: *Effect of number of speech segment pairs in learning speech representations for QbE speech search.*

search based on these feature representations are tested using a single thread on a workstation equipped with an Intel Xeon E5-2680 @ 2.7GHz CPU. The results show that the acoustic word embeddings outperform the frame-level feature representations (about 4% relative improvement in both MAP and P@N), and they reduce run-time computation since no dynamic time warping is required in QbE speech search. This suggests that learning acoustic word embeddings with temporal context is effective in terms of both accuracy and computational efficiency. In addition, we also find that the acoustic word embeddings based on multi-lingual BNFs outperform those based on spectral features (e. g., mel-frequency cepstral coefficients). This demonstrates that multilingual knowledge from resource-rich languages is helpful to learn acoustic word embeddings for QbE speech search.

3.4. Effect of number of speech segment pairs

We also investigated how the number of speech segment pairs in learning speech representations would affect the performance for QbE speech search. We randomly selected subsets of N=[M, 10k, 100k, 250k, 500k] speech segment pairs, where M represents the minimum speech segment pairs in Set 1 and Set 2 respectively. The evaluation results are plotted in Fig. 4.

We observe that acoustic word embeddings derived from Set 1 and Set 2 consistently improve the search results as the number of speech segment pairs increases. When we have more than 10k speech segment pairs, the acoustic word embeddings of Set 2 consistently outperform those of Set 1. This suggests that we can train a better deep CNN for acoustic word embeddings using a larger vocabulary, and it is important to have sufficient speech segment pairs to learn acoustic word embeddings for QbE speech search. In addition, we also reported the results of frame-level feature representations trained in both datasets. From Fig. 4 we can find that our proposed acoustic word embeddings can consistently give higher search accuracies than the frame-level feature representations when more than 100k speech segment pairs are available.

4. Conclusion

We have proposed a novel approach to learn convolutional neural acoustic word embeddings trained with temporal context padding for QbE speech search. The temporal context padding reduces the mismatch between the learning of embeddings and the use of embeddings on search content. Our proposed acoustic word embeddings can outperform the state-of-the-art frame-level feature representations and reduce run-time computation since no dynamic time warping is required in QbE speech search. Sufficient speech segment pairs with sufficient vocabulary coverage are important to learn acoustic word embeddings for QbE speech search. In the future, the discovery and selection of speech segment pairs will be worth exploring, and we will investigate recurrent neural networks, which are capable of modeling sequences, to obtain better acoustic word embeddings for this task.

5. Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61571363) and the China Scholarship Council (Grant No. 201706290169).

6. References

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

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

[3] 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 *Proc. ICASSP*, 2012, pp. 5157–5160.

[4] 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.

[5] G. Mantena, S. Achanta, and K. Prahallad, "Query-by-example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 5, pp. 946–955, 2014.

[6] H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Unsupervised bottleneck features for low-resource query-by-example spoken term detection," in *Proc. INTERSPEECH*, 2016, pp. 923–927.

[7] J. Tejedor *et al.*, "Comparison of methods for language-dependent and language-independent query-by-example spoken term detection," *ACM Trans. Inf. Syst.*, vol. 30, no. 3, p. 18, 2012.

[8] 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.

[9] C.-C. Leung *et al.*, "Toward high-performance language-independent query-by-example spoken term detection for mediaeval 2015: Post-evaluation analysis," in *Proc. INTERSPEECH*, 2016, pp. 3703–3707.

[10] Y. Yuan, C.-C. Leung, L. Xie, H. Chen, B. Ma, and H. Li, "Pairwise learning using multi-lingual bottleneck features for low-resource query-by-example spoken term detection," in *Proc. ICASSP*, 2017, pp. 5645–5649.

[11] S. Bengio and G. Heigold, "Word embeddings for speech recognition," in *Proc. INTERSPEECH*, 2014, pp. 1053–1057.

[12] K. Levin, K. Henry, A. Jansen, and K. Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings," in *Proc. ASRU*, 2013, pp. 410–415.

[13] H. Kamper, W. Wang, and K. Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," in *Proc. ICASSP*, 2016, pp. 4950–4954.

[14] S. Settle and K. Livescu, "Discriminative acoustic word embeddings: Tecurrent neural network-based approaches," in *Proc. SLT*, 2016, pp. 503–510.

[15] K. Levin, A. Jansen, and B. Van Durme, "Segmental acoustic indexing for zero resource keyword search," in *Proc. ICASSP*, 2015, pp. 5828–5832.

[16] S. Settle, K. Levin, H. Kamper, and K. Livescu, "Query-by-example search with discriminative neural acoustic word embeddings," in *Proc. INTERSPEECH*, 2017, pp. 2874–2878.

[17] Y. Yuan, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Learning neural network representation using cross-lingual bottleneck features with word-pair information," in *Proc. INTERSPEECH*, 2016, pp. 788–792.

[18] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2015, pp. 815–823.

[19] A. Jansen, S. Thomas, and H. Hermansky, "Weak top-down constraints for unsupervised acoustic model training," in *Proc. ICASSP*, 2013, pp. 8091–8095.

[20] H. Kamper, M. Elsner, A. Jansen, and S. Goldwater, "Unsupervised neural network based feature extraction using weak top-down constraints," in *Proc. ICASSP*, 2015, pp. 5818–5822.

[21] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: A cpu and gpu math compiler in python," in *Proc. 9th Python in Science Conf*, 2010, pp. 1–7.

[22] M. D. Zeiler, "Adadelta: an adaptive learning rate method," in *arXiv preprint arXiv:1212.5701*, 2012.